Lecture Notes in Computer Science 14959

The series Lecture Notes in Computer Science (LNCS), including its subseries Lecture Notes in Artificial Intelligence (LNAI) and Lecture Notes in Bioinformatics (LNBI), has established itself as a medium for the publication of new developments in computer science and information technology research, teaching, and education.

LNCS enjoys close cooperation with the computer science R & D community, the series counts many renowned academics among its volume editors and paper authors, and collaborates with prestigious societies. Its mission is to serve this international community by providing an invaluable service, mainly focused on the publication of conference and workshop proceedings and postproceedings. LNCS commenced publication in 1973.

Lorraine Goeuriot · Philippe Mulhem ·
Georges Quénot · Didier Schwab ·
Giorgio Maria Di Nunzio · Laure Soulier ·
Petra Galuščáková ·
Alba García Seco de Herrera ·
Guglielmo Faggioli · Nicola Ferro
Editors

Experimental IR Meets Multilinguality, Multimodality, and Interaction

15th International Conference of the CLEF Association, CLEF 2024
Grenoble, France, September 9–12, 2024
Proceedings, Part II

Springer

Editors

Lorraine Goeuriot
Université Grenoble Alpes, CNRS
Grenoble, France

Philippe Mulhem
Université Grenoble Alpes, CNRS
Grenoble, France

Georges Quénot
Université Grenoble Alpes, CNRS
Grenoble, France

Didier Schwab
Université Grenoble Alpes, CNRS
Grenoble, France

Giorgio Maria Di Nunzio
University of Padova
Padua, Italy

Laure Soulier
Sorbonne University
Paris, France

Petra Galuščáková
University of Stavanger
Stavanger, Norway

Alba García Seco de Herrera
University of Essex
Colchester, UK

Guglielmo Faggioli
University of Padova
Padua, Italy

Nicola Ferro
University of Padova
Padua, Italy

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-031-71907-3 ISBN 978-3-031-71908-0 (eBook)
https://doi.org/10.1007/978-3-031-71908-0

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Since 2000, the *Conference and Labs of the Evaluation Forum* (CLEF) has played a leading role in stimulating research and innovation in the domain of multimodal and multilingual information access. Initially founded as the *Cross-Language Evaluation Forum* and running in conjunction with the *European Conference on Digital Libraries* (ECDL/TPDL), CLEF became a standalone event in 2010 combining a peer-reviewed conference with a multi-track evaluation forum. The combination of the scientific program and the track-based evaluations at the CLEF conference creates a unique platform to explore information access from different perspectives, in any modality and language.

The CLEF conference has a clear focus on experimental information retrieval (IR) as seen in evaluation forums (like the CLEF Labs, TREC, NTCIR, FIRE, MediaEval, RomIP, TAC) with special attention to the challenges of multimodality, multilinguality, and interactive search, ranging from unstructured to semi-structured and structured data. The CLEF conference invites submissions on new insights demonstrated by the use of innovative IR evaluation tasks or in the analysis of IR test collections and evaluation measures, as well as on concrete proposals to push the boundaries of the Cranfield/TREC/CLEF paradigm.

CLEF 2024[1] was organized by the University of Grenoble Alpes, Grenoble, France, from 9 to 12 September 2024. CLEF 2024 was the 15th year of the CLEF Conference and the 25th year of the CLEF initiative as a forum for IR Evaluation, so it marked an important anniversary for CLEF. The conference format remained the same as in past years and consisted of keynotes, contributed papers, lab sessions, and poster sessions, including reports from other benchmarking initiatives from around the world. All sessions were organized in presence but also allowing for remote participation for those who were not able to attend physically. The CLEF 25th anniversary paper, a kind of *una tantum* paper to celebrate the event, was reviewed by one of the Program Chairs. As usual, the lab overview papers were reviewed by the Lab Chairs.

CLEF 2024 continued the initiative introduced in the 2019 edition, during which the *European Conference on Information Retrieval (ECIR)* and CLEF joined forces: ECIR 2023[2] hosted a special session dedicated to CLEF Labs where lab organizers presented the major outcomes of their Labs and their plans for ongoing activities, followed by a poster session to favour discussion during the conference. This was reflected in the ECIR 2024 proceedings, where CLEF Lab activities and results were reported as short papers. The goal was not only to engage the ECIR community in CLEF activities but also to disseminate the research results achieved during CLEF evaluation cycles as papers submitted to ECIR.

The following scholars were invited to give a keynote talk at CLEF 2024: Paula Carvalho (INESC-ID, Lisboa, Portugal) and Aurélie Névéol (Université Paris-Saclay, LISN, CNRS, France).

1 https://clef2024.clef-initiative.eu/
2 https://ecir2024.org/

CLEF 2024 received a total of 25 scientific submissions, of which a total of 11 papers (7 long, 3 short & 1 position) were accepted. Each submission was reviewed in double-blind fashion by at least two program committee members, and the program chairs oversaw the reviewing and follow-up discussions. Several papers were a product of international collaboration. This year, researchers addressed the following important challenges in the community: factual reporting and political bias; sexism, discrimination, and misinformation; information retrieval and recommendation; information retrieval for decision making; document sanitization for information release and retrieval; evaluation dataset for knowledge acquisition; evaluation with gen-IR; medical entity linking; and classification with large language models.

Like in previous editions, since 2015, CLEF 2024 continued inviting CLEF lab organizers to nominate a "best of the labs" paper, among those submitted in the CLEF 2023 labs, that was reviewed as a full paper submission to the CLEF 2024 conference, according to the same review criteria and PC. 6 full papers were accepted for this "best of the labs" section.

The conference integrated a series of workshops presenting the results of lab-based comparative evaluations. A total of 23 lab proposals were received and evaluated in peer review based on their innovation potential and the quality of the resources created. The 14 selected labs represented scientific challenges based on new datasets and real-world problems in multimodal and multilingual information access. These datasets provide unique opportunities for scientists to explore collections, to develop solutions for these problems, to receive feedback on the performance of their solutions, and to discuss the challenges with peers at the workshops. In addition to these workshops, the labs reported results of their year-long activities in overview talks and lab sessions. Overview papers describing each of the labs are provided in this volume. The full details for each lab are contained in a separate publication, the Working Notes[3].

The 14 labs running as part of CLEF 2024 comprised mainly labs that continued from previous editions at CLEF (BioASQ, CheckThat!, eRisk, EXIST, iDPP, ImageCLEF, JOKER, LifeCLEF, LongEval, PAN, SimpleText, and Touché) and new pilot/workshop activities (ELOQUENT and qCLEF). In the following we give a few details for each of the labs organized at CLEF 2024 (presented in alphabetical order):

BioASQ: Large-scale biomedical semantic indexing and question answering[4] aimed to push the research frontier towards systems that use the diverse and voluminous information available online to respond directly to the information needs of biomedical scientists. It offered the following tasks. *Task 1 - b: Biomedical Semantic Question Answering*: benchmark datasets of biomedical questions, in English, along with gold standard (reference) answers constructed by a team of biomedical experts. The participants had to respond with relevant articles, and snippets from designated resources, as well as exact and "ideal" answers. *Task 2 - Synergy: Question Answering for developing problems*: biomedical experts posed unanswered questions for developing problems, such as COVID-19, received the responses provided by the participating systems, and provided feedback, together with updated questions in an iterative procedure that aimed

[3] Faggioli, G., Ferro, N., Galuščáková, P., and García Seco de Herrera, A. editors (2024). *CLEF 2024 Working Notes*. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073.

[4] http://www.bioasq.org/workshop2024

to facilitate the incremental understanding of developing problems in biomedicine and public health. *Task 3 - MultiCardioNER: Multiple clinical entity detection in multilingual medical content*: focused on the automatic detection and normalization of mentions of four clinical entity types, namely diseases, symptoms, procedures, and medications, in cardiology clinical case documents in Spanish, English, Italian, and Dutch. *BioNNE: Nested NER in Russian and English*: dealt with nested named-entity recognition (NER) in PubMed abstracts in Russian and English. The train/dev datasets included annotated mentions of disorders, anatomical structures, chemicals, diagnostic procedures, and biological functions. Participants were encouraged to apply cross-language (Russian to English) and cross-domain techniques.

CheckThat! Lab on Checkworthiness, Subjectivity, Persuasion, Roles, Authorities and Adversarial Robustness[5] provided a diverse collection of challenges to the research community interested in developing technology to support and understand the journalistic verification process. The tasks went from core verification tasks such as assessing the check-worthiness of a text to understanding the strategies used to influence the audience and identifying the stance of relevant characters on questionable affairs. It offered the following tasks. *Task 1 - Check-worthiness estimation*: asked to assess whether a statement, sourced from either a tweet or a political debate, warrants fact-checking. *Task 2 - Subjectivity*: given a sentence from a news article, it asks to determine whether it is subjective or objective. *Task 3 - Persuasion Techniques*: given a news article and a list of 23 persuasion techniques organized into a 2-tier taxonomy, including logical fallacies and emotional manipulation techniques that might be used to support flawed argumentation, it asked to identify the spans of texts in which each technique occurs. *Task 4 - Detecting hero, villain, and victim from memes*: asked to determine the roles of entities within memes, categorizing them as “hero”, “villain”, “victim”, or “other” through a multi-class classification approach that considers the systematic modelling of multimodal semiotics. *Task 5 - Authority Evidence for Rumor Verification*: given a rumor expressed in a tweet and a set of authorities for that rumor, it asked to retrieve up to 5 evidence tweets from the authorities’ timelines, and determine whether the rumor is supported, refuted, or unverifiable according to the evidence. *Task 6 - Robustness of Credibility Assessment with Adversarial Examples*: the task was realised in five domains: style-based news bias assessment (HN), propaganda detection (PR), fact checking (FC), rumour detection (RD), and COVID-19 misinformation detection (C19). For each domain, the participants were provided with three victim models, trained for the corresponding binary classification task, as well as a collection of 400 text fragments. Their aim was to prepare adversarial examples which preserve the meaning of the original examples, but were labelled differently by the classifiers.

ELOQUENT shared tasks for evaluation of generative language model quality[6] provided a set of tasks for evaluating the quality of generative language models. It offered the following tasks. *Task 1 - Topical competence*: tested and verified a model’s understanding of an application domain and specific topic of interest. *Task 2 - Veracity and hallucination*: tested how the truthfulness or veracity of automatically generated text can be assessed. *Task 3 - Robustness*: tested the capability of a model to handle input

[5] http://checkthat.gitlab.io/

[6] https://eloquent-lab.github.io/

variation – e.g., dialectal, sociolectal, and cross-cultural – as represented by a set of equivalent but non-identical varieties of input prompts. *Task 4 - Voight Kampff*: explored whether automatically generated text can be distinguished from human-authored text. This task was organized in collaboration with the PAN lab at CLEF.

eRisk: Early Risk Prediction on the Internet[7] explored the evaluation methodology, effectiveness metrics, and practical applications (particularly those related to health and safety) of early risk detection on the Internet. It offered the following tasks. *Task 1 - Search for symptoms of depression*: consisted of ranking sentences from a collection of user writings according to their relevance to a depression symptom. The participants had to provide rankings for the 21 symptoms of depression from the BDI Questionnaire. *Task 2 - Early Detection of Signs of Anorexia*: consisted in performing a task on early risk detection of anorexia. The challenge consisted of sequentially processing pieces of evidence to detect early traces of anorexia as soon as possible. *Task 3 - Measuring the severity of the signs of Eating Disorders*: consisted of estimating the level of features associated with a diagnosis of eating disorders from a thread of user submissions. For each user, the participants were given a history of postings and the participants had to fill in a standard eating disorder questionnaire.

EXIST: sEXism Identification in Social neTworks[8] aimed to capture and categorize sexism, from explicit misogyny to other subtle behaviours, in social networks. Participants were asked to classify tweets in English and Spanish according to the type of sexism they enclose and the intention of the persons that wrote the tweets. It offered the following tasks. *Task 1 - Sexism Identification in Tweets*: was a binary classification. The systems had to decide whether or not a given tweet contains sexist expressions or behaviours (i.e., it is sexist itself, describes a sexist situation, or criticizes a sexist behaviour). *Task 2 - Source Intention in Tweets*: aimed to categorize the message according to the intention of the author, which provides insights in the role played by social networks on the emission and dissemination of sexist messages. *Task 3 - Sexism Categorization in Tweets*: many facets of a woman's life may be the focus of sexist attitudes including domestic and parenting roles, career opportunities, sexual image, and life expectations, to name a few. Automatically detecting which of these facets of women are being more frequently attacked in social networks will facilitate the development of policies to fight against sexism. *Task 4 - Sexism Identification in Memes*: was a binary classification task consisting of deciding whether or not a given meme is sexist. *Task 5 - Source Intention in Memes*: aimed to categorize the meme according to the intention of the author, which provides insights in the role played by social networks in the emission and dissemination of sexist messages.

iDPP: Intelligent Disease Progression Prediction[9] Amyotrophic Lateral Sclerosis (ALS) and Multiple Sclerosis (MS) are chronic diseases characterized by progressive or alternate impairment of neurological functions (motor, sensory, visual, cognitive). Patients have to manage alternating periods in hospital with care at home, experiencing a constant uncertainty regarding the timing of the disease acute phases and facing a considerable psychological and economic burden that also involves their caregivers.

---
7 https://erisk.irlab.org/
8 http://nlp.uned.es/exist2024/
9 https://brainteaser.health/open-evaluation-challenges/idpp-2024/

Clinicians, on the other hand, need tools able to support them in all the phases of the patient treatment, to suggest personalized therapeutic decisions, and to indicate urgently needed interventions. It offered the following tasks. *Task 1 – Predicting ALSFRS-R score from sensor data (ALS)*: focused on predicting the ALSFRS-R score (ALS Functional Rating Scale - Revised), assigned by medical doctors roughly every three months, from the sensor data collected via the app. The ALSFRS-R score is a somehow "subjective" evaluation performed by a medical doctor and this task will help in answering a currently open question in the research community, i.e., whether it could be derived from objective factors. *Task 2 – Predicting patient self-assessment score from sensor (ALS)*: focused on predicting the self-assessment score assigned by patients from the sensor data collected via the app. If the self-assessment performed by patients, more frequently than the assessment performed by medical doctors every three months or so, can be reliably predicted by sensor and app data, we can imagine a proactive application which, monitoring the sensor data, alerts the patient if an assessment is needed. *Task 3 – Predicting relapses from EDDS sub-scores and environmental data (MS))*: focused on predicting a relapse using environmental data and EDSS (Expanded Disability Status Scale) sub-scores. This task will allow us to assess whether exposure to different pollutants is a useful variable in predicting a relapse.

ImageCLEF: Multimedia Retrieval[10]aimed at evaluating the technologies for annotation, indexing, classification, and retrieval of multimodal data. Its main objective resided in providing access to large collections of multimodal data for multiple usage scenarios and domains. Considering the experience of the last four successful editions, ImageCLEF 2024 continued to address a diversity of applications, namely medical, social media, and Internet, and recommending, giving to the participants the opportunity to deal with interdisciplinary approaches and domains. It offered the following tasks. *Task 1 - ImageCLEFmedical*: continued the tradition of bringing together several initiatives for medical applications fostering cross-exchanges, namely: (i) caption task with medical concept detection and caption prediction, (ii) GAN task on synthetic medical images generated with GANs, (iii) MEDVQA-GI task for medical images generation based on text input, and (iv) Mediqa task with a new use-case on multimodal dermatology response generation. *Task 2 - Image Retrieval/Generation for Arguments*: given a set of arguments, asked to return for each argument several images that help to convey the argument's premise, that is, suitable images to depict what is described in the argument. *Task 3 - ImageCLEFrecommending*: focused on content recommendation for cultural heritage content. Despite current advances in content-based recommendation systems, there is limited understanding of how well these perform and how relevant they are for the final end-users. This task aimed to fill this gap by benchmarking different recommendation systems and methods. *Task 4 - ImageCLEFtoPicto*: aimed to provide a translation in pictograms from a natural language, either from (i) text or (ii) speech understandable by the users, in this case, people with language impairments, as pictogram generation is an emerging and significant domain in natural language processing, with multiple potential applications, enabling communication with individuals who have disabilities, aiding in medical settings for individuals who do not speak the language of a country, and also enhancing user understanding in the service industry..

[10] https://www.imageclef.org/2024

JOKER: Automatic Humour Analysis[11] aimed to foster research on automated processing of verbal humour, including tasks such as retrieval, classification, interpretation, generation, and translation. It offered the following tasks. *Task 1 - Humour-aware information retrieval*: aimed at retrieving short humorous texts from a document collection. *Task 2 - Humour classification according to genre and technique*: aimed at classifying short texts of humour among the different classes such as Irony, Sarcasm, Exaggeration, Incongruity, Absurdity, etc. *Task 3 - Pun translation*: aimed to translate English punning jokes into French preserving wordplay form and wordplay meaning.

LifeCLEF: species identification and prediction[12] was dedicated to the large-scale evaluation of biodiversity identification and prediction methods based on artificial intelligence. It offered the following tasks. *Task 1 - BirdCLEF*: bird species recognition in audio soundscapes. *Task 2 - FungiCLEF*: fungi recognition from images and metadata. *Task 3 - GeoLifeCLEF*: remote sensing-based prediction of species. *Task 4 - PlantCLEF*: global-scale plant identification from images. *Task 5 - SnakeCLEF*: snake species identification in medically important scenarios.

LongEval: Longitudinal Evaluation of Model Performance[13] focused on evaluating the temporal persistence of information retrieval systems and text classifiers. The goal was to develop temporal information retrieval systems and longitudinal text classifiers that survive through dynamic temporal text changes, introducing time as a new dimension for ranking models' performance. It offered the following tasks. *Task 1 - LongEval-Retrieval*: aimed to propose a temporal information retrieval system which can handle changes over time. The proposed retrieval system should demonstrate temporal persistence on Web documents. This task had 2 sub-tasks focusing on short-term and long-term persistence. *Task 2 - LongEval-Classification* aimed to propose a temporal persistence classifier which can mitigate performance drop over short and long periods of time compared to a test set from the same time frame as training. This task had 2 sub-tasks focusing on short-term and long-term persistence.

PAN: Digital Text Forensics and Stylometry[14] aimed to advance the state of the art and provide for an objective evaluation on newly developed benchmark datasets in those areas. It offered the following tasks. *Task 1 - Multi-Author Writing Style Analysis*: given an English document, asked to determine at which paragraphs the author changes. Examples varied in difficulty from easy to hard depending on the topical homogeneity of the paragraphs. *Task 2 - Multilingual Text Detoxification*: given a toxic piece of text, asked to re-write it in a non-toxic way while saving the main content as much as possible. Texts were provided in 7 languages. *Task 3 - Oppositional Thinking Analysis*: given an English or Spanish online message, asked to determine whether it is a conspiracy theory or critical thinking. In former case, find the core elements of the conspiracy narrative. *Task 4 - Generative AI Authorship Verification*: given a document, asked to determine whether the author is a human or a language model. In collaboration with the ELOQUENT lab.

[11] http://joker-project.com/
[12] http://www.lifeclef.org/
[13] https://clef-longeval.github.io/
[14] http://pan.webis.de/

qCLEF: QuantumCLEF[15] Quantum Computing (QC) is a rapidly growing field, involving an increasing number of researchers and practitioners from different backgrounds who develop new methods that leverage quantum computers to perform faster computations. QuantumCLEF provided an evaluation infrastructure to design and develop QC algorithms and, in particular, for Quantum Annealing (QA) algorithms, for Information Retrieval and Recommender Systems. It offered the following tasks. *Task 1 - Feature Selection*: focused on applying quantum annealers to find the most relevant subset of features to train a learning model, e.g., for ranking. This problem is very impactful, since many IR and RS systems involve the optimization of learning models, and reducing the dimensionality of the input data can improve their performance. *Task 2 - Clustering*: focused on using quantum annealing to cluster different documents in the form of embeddings to ease the browsing process of large collections. Clustering can be helpful for organizing large collections, helping users to explore a collection and providing similar search results to a given query. Furthermore, it can be helpful to divide users according to their interests or build user models with the cluster centroids speeding up the runtime of the system or its effectiveness for users with limited data. Clustering is however a very complex task in the case of QA since it is possible to perform clustering only considering a limited number of items and clusters due to the architecture of quantum annealers. A baseline using K-medoids clustering with cosine distance was used as an overall alternative.

SimpleText: Improving Access to Scientific Texts for Everyone[16] addressed technical and evaluation challenges associated with making scientific information accessible to a wide audience, students, and experts. Appropriate reusable data and benchmarks were provided for scientific text summarization and simplification. *Task 1 - Retrieving passages to include in a simplified summary*: given a popular science article targeted to a general audience, aimed at retrieving passages which can help to understand this article, from a large corpus of academic abstracts and bibliographic metadata. Relevant passages should relate to any of the topics in the source article. *Task 2 - Identifying and explaining difficult concepts*: aimed to decide which concepts in scientific abstracts require explanation and contextualization in order to help a reader understand the scientific text. *Task 3 - Simplify Scientific Text*: aimed to provide a simplified version of sentences extracted from scientific abstracts. Participants were provided with popular science articles and queries and matching abstracts of scientific papers, split into individual sentences. *Task 4 - Tracking the State-of-the-Art in Scholarly Publications*: aimed to develop systems which, given the full text of an AI paper, are capable of recognizing whether an incoming AI paper indeed reports model scores on benchmark datasets, and if so, to extract all pertinent (Task, Dataset, Metric, Score) tuples presented within the paper.

Touché: Argumentation Systems[17] aimed foster the development of technologies that support people in decision-making and opinion-forming and to improve our understanding of these processes. It offered the following tasks. *Task 1 - Human Value Detection*: given a text, for each sentence, asked to detect which human values the sentence refers to and whether this reference (partially) attains or (partially) constrains the value.

15 https://qclef.dei.unipd.it/
16 http://simpletext-project.com/
17 https://touche.webis.de/

Task 2 - Ideology and Power Identification in Parliamentary Debates: given a parliamentary speech in one of several languages, asked to identify the ideology of the speaker's party and identify whether the speaker's party is currently governing or in opposition.
Task 3 - Image Retrieval for Arguments: given an argument, asked to retrieve or generate images that help to convey the argument's premise.

The success of CLEF 2024 would not have been possible without the huge effort of several people and organizations, including the CLEF Association[18], the Program Committee, the Lab Organizing Committee, the reviewers, and the many students and volunteers who contributed.

We thank the Friends of SIGIR program for covering the registration fees for a number of student delegates.

July 2024

Lorraine Goeuriot
Philippe Mulhem
Georges Quénot
Didier Schwab
Giorgio Maria Di Nunzio
Laure Soulier
Petra Galuščáková
Alba García Seco de Herrera
Guglielmo Faggioli
Nicola Ferro

[18] https://www.clef-initiative.eu/#association

Organization

General Chairs

Lorraine Goeuriot	Université Grenoble Alpes, France
Philippe Mulhem	Université Grenoble Alpes, France
Georges Quénot	Université Grenoble Alpes, France
Didier Schwab	Université Grenoble Alpes, France

Program Chairs

Giorgio Maria Di Nunzio	University of Padua, Italy
Laure Soulier	Sorbonne Université, France

Lab Chairs

Petra Galuščáková	University of Stavanger, Norway
Alba García Seco de Herrera	University of Essex, UK

Lab Mentorship Chair

Liana Ermakova	Université de Bretagne Occidentale, France
Florina Piroi	TU Wien, Austria

Proceedings Chairs

Guglielmo Faggioli	University of Padua, Italy
Nicola Ferro	University of Padua, Italy

CLEF Steering Committee

Steering Committee Chair

Nicola Ferro	University of Padua, Italy

Deputy Steering Committee Chair for the Conference

Paolo Rosso	Universitat Politècnica de València, Spain

Deputy Steering Committee Chair for the Evaluation Labs

Martin Braschler	Zurich University of Applied Sciences, Switzerland

Members

Avi Arampatzis	Democritus University of Thrace, Greece
Alberto Barrón-Cedeño	University of Bologna, Italy
Khalid Choukri	Evaluations and Language resources Distribution Agency, France
Fabio Crestani	Università della Svizzera italiana, Switzerland
Carsten Eickhoff	University of Tübingen, Germany
Norbert Fuhr	University of Duisburg-Essen, Germany
Anastasia Giachanou	Utrecht University, The Netherlands
Lorraine Goeuriot	Université Grenoble Alpes, France
Julio Gonzalo	National Distance Education University (UNED), Spain
Donna Harman	National Institute for Standards and Technology, USA
Bogdan Ionescu	University “Politehnica” of Bucharest, Romania
Evangelos Kanoulas	University of Amsterdam, The Netherlands
Birger Larsen	University of Aalborg, Denmark
David E. Losada	Universidade de Santiago de Compostela, Spain
Mihai Lupu	Vienna University of Technology, Austria
Maria Maistro	University of Copenhagen, Denmark
Josiane Mothe	IRIT, Université de Toulouse, France
Henning Müller	University of Applied Sciences Western Switzerland (HES-SO), Switzerland

Jian-Yun Nie	Université de Montréal, Canada
Gabriella Pasi	University of Milano-Bicocca, Italy
Eric SanJuan	University of Avignon, France
Giuseppe Santucci	Sapienza University of Rome, Italy
Laure Soulier	Pierre and Marie Curie University (Paris 6), France
Theodora Tsikrika	Information Technologies Institute, Centre for Research and Technology Hellas (CERTH), Greece
Christa Womser-Hacker	University of Hildesheim, Germany

Past Members

Paul Clough	University of Sheffield, UK
Djoerd Hiemstra	Radboud University, The Netherlands
Jaana Kekäläinen	University of Tampere, Finland
Séamus Lawless	Trinity College Dublin, Ireland
Carol Peters	ISTI, National Council of Research (CNR), Italy
Emanuele Pianta	Centre for the Evaluation of Language and Communication Technologies, Italy
Maarten de Rijke	University of Amsterdam, The Netherlands
Jacques Savoy	University of Neuchâtel, Switzerland
Alan Smeaton	Dublin City University, Ireland

Supporters and Sponsors

SIGIR
Special Interest Group
on Information Retrieval

Contents – Part II

Contents – Part I

Lab Overviews

Overview of BioASQ 2024: The Twelfth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering

Anastasios Nentidis[1,2](✉), Georgios Katsimpras[1], Anastasia Krithara[1], Salvador Lima-López[3], Eulàlia Farré-Maduell[3], Martin Krallinger[3], Natalia Loukachevitch[4], Vera Davydova[5], Elena Tutubalina[5,6,7], and Georgios Paliouras[1]

[1] National Center for Scientific Research "Demokritos", Athens, Greece
{tasosnent,gkatsibras,akrithara,paliourg}@iit.demokritos.gr
[2] Aristotle University of Thessaloniki, Thessaloniki, Greece
[3] Barcelona Supercomputing Center, Barcelona, Spain
{slimalop,efarre,martin.krallinger}@bsc.es
[4] Moscow State University, Moscow, Russia
[5] Sber AI, Moscow, Russia
[6] Artificial Intelligence Research Institute, Moscow, Russia
[7] Kazan Federal University, Kazan, Russia

Abstract. This is an overview of the twelfth edition of the BioASQ challenge in the context of the Conference and Labs of the Evaluation Forum (CLEF) 2024. BioASQ is a series of international challenges promoting advances in large-scale biomedical semantic indexing and question answering. This year, BioASQ consisted of new editions of the two established tasks b and Synergy, and two new tasks: a) MultiCardioNER on the adaptation of clinical entity detection to the cardiology domain in a multilingual setting, and b) BIONNE on nested NER in Russian and English. In this edition of BioASQ, 37 competing teams participated with more than 700 distinct submissions in total for the four different shared tasks of the challenge. Similarly to previous editions, most of the participating systems achieved competitive performance, suggesting the continuous advancement of the state-of-the-art in the field.

Keywords: Biomedical knowledge · Semantic Indexing · Question Answering

1 Introduction

The BioASQ challenge has been focusing on the advancement of the state-of-the-art in large-scale biomedical semantic indexing and question answering (QA) for more than 10 years [50]. To this end, it organizes different shared tasks annually,

L. Goeuriot et al. (Eds.): CLEF 2024, LNCS 14959, pp. 3–27, 2024.
https://doi.org/10.1007/978-3-031-71908-0_1

developing respective benchmark datasets that represent the real information needs of experts in the biomedical domain. This allows the participating teams from around the world, who work on the development of systems for biomedical semantic indexing and question answering, to benefit from the publicly available datasets, evaluation infrastructure, and exchange of ideas in the context of the BioASQ challenge and workshop.

Here, we present the shared tasks and the datasets of the twelfth BioASQ challenge in 2024, as well as an overview of the participating systems and their performance. The remainder of this paper is organized as follows. First, Sect. 2 presents a general description of the shared tasks, which took place from January to May 2024, and the corresponding datasets developed for the challenge. Then, Sect. 3 provides a brief overview of the participating systems for the different tasks. Detailed descriptions for some of the systems are available in the proceedings of the lab. Subsequently, in Sect. 4, we present the performance of the systems for each task, based on state-of-the-art evaluation measures or manual assessment. Finally, in Sect. 5 we draw some conclusions regarding the 2024 edition of the BioASQ challenge.

2 Overview of the Tasks

The twelfth edition of the BioASQ challenge consisted of four tasks: (1) a biomedical question answering task (task b), (2) a task on biomedical question answering for open developing issues (task Synergy), both tasks considering documents in English, (3) a new task focused on the automatic detection of disease and drug mentions (task MultiCardioNER), considering cardiology clinical case documents in Spanish, English, and Italian, and (4) a new task on NLP challenges on biomedical nested named entity recognition (NER) systems for English and Russian languages (task BIONNE) [43]. In this section, we first describe this year's editions of the two established tasks b (task 12b) and Synergy (Synergy 12) [37] with a focus on differences from previous editions of the challenge [36,39]. Additionally, we also present the new tasks MultiCardioNER on multiple clinical entity detection in multilingual medical content [31], and BIONNE on nested NER in Russian and English [12].

2.1 Task 12b

The twelfth edition of task b (task 12b) focuses on a large-scale question-answering scenario in which the participants are required to develop systems for all the stages of biomedical question answering. As in previous editions, the task examines four types of questions: "yes/no", "factoid", "list" and "summary" questions [6]. In this edition, the training dataset provided to the participating teams for the development of their systems consisted of 5,049 biomedical questions from previous versions of the challenge annotated with ground-truth relevant material, that is, articles, snippets, and answers [24,37]. Table 1 shows the details of both training and test datasets for task 12b. The test data for task

12b were split into four independent bi-weekly batches consisting of 85 questions each, as presented in Table 1.

Table 1. Statistics on the training and test datasets of task 12b. The numbers for the documents and snippets refer to averages per question.

Batch	Size	Yes/No	List	Factoid	Summary	Documents	Snippets
Train	5,049	1,357	967	1,515	1,210	9.06	11.91
Test 1	85	25	21	21	18	3.20	4.36
Test 2	85	26	18	19	22	2.72	3.69
Test 3	85	24	19	26	16	2.45	3.36
Test 4	85	27	22	19	17	2.18	3.44
Total	5,389	1,459	1,047	1,600	1,283	8.65	11.4

As in previous editions of task b, task 12b was also divided into distinct phases. Contrary to previous editions, however, task 12b was divided into three phases: a) In phase A, a test set consisting of the bodies of biomedical questions, written in English, was released and the participants had 24 h to identify and submit relevant PubMed/MEDLINE-article abstracts, and snippets extracted from them. b) In phase A+, which runs in parallel to phase A, the participants could also submit *exact answers*, that is entity names or short phrases, and *ideal answers*, that is, natural language summaries of the requested information for the same questions. c) In phase B, which runs after the completion of phases A and A+, some relevant articles and snippets were released for these questions, and the participating systems had another 24 h to respond with *exact* and *ideal answers* taking this manually selected material into account.

For example, for the "yes/no" question "Is levosimendan effective for amyotrophic lateral sclerosis?", the systems in phase A should respond with relevant documents and snippets useful for providing an answer. In parallel, systems participating in phase A+ could also attempt to provide the *exact* and *ideal answer*, which are "No." and "No. Levosimendan was not superior to placebo in maintaining respiratory function in a broad population with amyotrophic lateral sclerosis. Although levosimendan was generally well tolerated, increased heart rate and headache occurred more frequently with levosimendan than with placebo." respectively. In Phase B, sufficient relevant documents and snippets were also released, and the participating systems had to return *exact* and *ideal answer* again, exploiting this information.

2.2 Task Synergy 12

The task Synergy was introduced three years ago [38] envisioning a continuous dialog between the experts and the automated question-answering systems. In this model, the systems provide relevant material and answers to the experts who posed some open questions for developing problems. The experts assess these responses and feed their assessment back to the systems, including the information on whether the material retrieved is sufficient for providing an answer to this question ("answer ready"). This feedback is then exploited by the systems together with new material, that becomes available in the meantime, to provide updated responses to the experts. This process proceeds with new feedback and new responses for the same open questions that persist, in an iterative way, organized in rounds.

After twelve rounds in the context of BioASQ9 [25] and BioASQ10 [42], focusing on open questions about the COVID-19 pandemic, in BioASQ11 we extended the Synergy task with four rounds open to any developing problem [41]. In BioASQ12, we continue in this setting with four more bi-weekly rounds open to any developing problem of interest for the six biomedical experts who participated this year. As in previous versions, the open questions were not required to have definite answers, and a distinct version of PubMed/MEDLINE was designated per round for relevant material retrieval. A set of 311 questions with respective incremental expert feedback and answers from the previous versions of the task, was available as a development set. This year, 73 distinct questions were used in the new rounds of the task. Of them, 18 questions were persisting from the previous versions and 55 were new. The distribution of the Synergy 12 questions round is shown in Table 2.

Table 2. Distribution of the questions of Task Synergy 12 per round.

Round	Size	Yes/No	List	Factoid	Summary	Answer ready
1	72	11	29	17	15	33
2	72	11	29	18	14	46
3	64	10	24	16	14	50
4	64	10	24	17	13	57

The same types of questions (yes/no, factoid, list, and summary) and answers (*exact* and *ideal*) are examined in this task as in task 12b, and the same evaluation measures are adopted for system assessment. However, task Synergy is not structured into phases, with both relevant material and answers received together. For new questions, only relevant material was required until the expert marked a question as "answer ready". Then, both new relevant material and answers are expected for it in subsequent rounds. In case a question receives a definite answer that is not expected to change, the expert can mark it as "closed" to be excluded from the remaining rounds.

2.3 Task MultiCardioNER

The MultiCardioNER shared task is a continuation of previous tasks focused on the detection of named entities in clinical case reports in Spanish such as DisTEMIST [35] for diseases, MedProcNER/ProcTEMIST [30] for procedures and SympTEMIST [29] for signs and symptoms. While these previous tasks use a general collection of texts from multiple clinical specialties, MultiCardioNER focuses on the creation of systems that can detect diseases and drugs specifically in the cardiology domain. For this purpose, in addition to the DisTEMIST and DrugTEMIST corpus (which include disease and drug annotations, respectively, for the same collection of varied clinical case reports), participants were provided the CardioCCC corpus of cardiological documents. This new dataset includes 508 documents, out of which 250 were reserved to be used as test set and the rest were released so that participants could use them as they saw fit. As an added novelty, this task introduces a multilingual aspect with the release of the Gold Standard drug annotations in English and Italian, as well as in Spanish. Table 3 presents a summary of the different corpora used as part of the task.

Table 3. Statistics for the datasets provided for MultiCardioNER. "Annot." stands for "annotations", while "Chars" stands for "characters". Unique annotations refer to the number of distinct annotated strings after converting all annotations to lowercase. The number of tokens has been calculated using the following spaCy models: "es_core_news_sm", "en_core_web_sm" and "it_core_news_sm".

Dataset	Lang.	Entity	Docs	Tokens	Chars	Annot.	Unique Annot.
DisTEMIST	ES	Diseases	1,000	406,137	2,335,968	10,664	6,739
DrugTEMIST	ES	Drugs	1,000	406,137	2,335,968	2,778	925
	EN	Drugs	1,000	404,194	2,230,631	2,814	875
	IT	Drugs	1,000	421,251	2,393,002	2,808	893
CardioCCC	ES	Diseases	508	568,297	3,215,774	18,232	7,692
	ES	Drugs	508	568,297	3,215,774	4,227	755
	EN	Drugs	508	576,772	3,114,833	4,231	734
	IT	Drugs	508	595,332	3,345,466	4,385	752

All in all, MultiCardioNER is divided into two different subtracks:

- **Subtrack 1 (CardioDis)**. This track focuses on the adaptation of disease recognition systems to the cardiology domain in Spanish. Some examples of cardiology-specific diseases would be "atrial flutter with rapid ventricular response" or "Takotsubo syndrome". Participants were provided the DisTEMIST corpus [35], as well as a new collection of 258 cardiology-specific clinical case reports annotated with diseases (CardioCCC). They were allowed to distribute the data collections however they saw fit in order to achieve the best system possible. The evaluation was done in the second half of the CardioCCC corpus, made up of 250 documents, using strict, micro-averaged precision, recall and F1-score. The annotation for these documents was done following the DisTEMIST annotation guidelines, which are available on Zenodo[1]
- **Subtrack 2 (MultiDrug)**. This track focuses on the multilingual (Spanish, English and Italian) adaptation of medication recognition systems to the cardiology domain. Some examples of medication entities are "nytroglicerine" and "clopidogrel". For this track, participants were provided the

[1] https://zenodo.org/doi/10.5281/zenodo.6458078.

DrugTEMIST dataset, which is a companion corpus to the previously-released DisTEMIST, ProcTEMIST and SympTEMIST corpora that offer annotations of medications for the same collection of texts and hadn't been released until now. As in the previous track, a portion of the cardiology-specific dataset CardioCCC was released during the training phase and another was reserved to be used as test set. While the original versions of both datasets were created using Spanish texts, a machine-translated version in English and Italian was revised and annotated by clinical experts native in each language. These documents were annotated following the DrugTEMIST guidelines, published specifically for this task and available on Zenodo[2]. The evaluation for this subtrack was also done using strict, micro-averaged precision, recall and F1-score, with every language being evaluated separately.

The MultiCardioNER datasets are publicly available to download on Zenodo[3]. In addition to the Gold Standard datasets, a background set of related clinical case reports (including both cardiology and non-cardiology documents) was also released as part of the task. This background set includes 7,625 documents and is available in Spanish, English and Italian. There are some documents originally being written in each language, with the rest being translated via machine translation. A Silver Standard that aggregates the predictions of the participant systems for this documents will also be released and uploaded to the same repository.

MultiCardioNER is promoted by Spanish and European projects such as DataTools4Heart, AI4HF, BARITONE and AI4ProfHealth and organized by the Barcelona Supercomputing Center (BSC) in collaboration with BioASQ. A more in-depth analysis of the MultiCardioNER Gold Standard, guidelines and additional resources is presented in the MultiCardioNER overview paper [31].

2.4 Task BioNNE

Given that most biomedical datasets and named entity recognition (NER) methods are designed to identify flat, non-nested mention structures, we introduce the Biomedical Nested Named Entity Recognition (BioNNE) shared task this year. For instance, in the text "[[[eye] movement] disorders]", nested annotations assist in identifying and distinguishing both the broader categories, like medical conditions and physiological functions, as well as the specific anatomical parts involved. The main task focuses on extracting and classifying biomedical nested named entities from unstructured PubMed abstracts available in both Russian and English. It is divided into three tracks:

- **Bilingual**: Participants develop a single multilingual NER model using data in both Russian and English, generating predictions for each language.
- **English-oriented or Russian-oriented**: Participants build a nested NER model specifically for abstracts in one target language, either English or Russian.

[2] https://zenodo.org/doi/10.5281/zenodo.11065432.
[3] https://zenodo.org/doi/10.5281/zenodo.10948354.

The training and validation sets for the BioNNE competition were derived from a subset of the NEREL-BIO dataset [34]. This dataset enhances the original NEREL [33] dataset, which was designed for the general domain by incorporating biomedical entity types. We made improvements by correcting annotator errors, merging the `PRODUCT` and `DEVICE` classes into a unified `DEVICE` class, and selecting the eight most frequent medical entities: `FINDING`, `DISO`, `INJURY_POISONING`, `PHYS`, `DEVICE`, `LABPROC`, `ANATOMY`, and `CHEM`. The final dataset includes 662 annotated PubMed abstracts in Russian and 104 parallel abstracts in both Russian and English. In total, there are 40,782 annotated entities in Russian and 8,099 in English. A new test set was developed specifically for the shared task, including 154 abstracts in both English and Russian, each containing ≈10k annotated entities.

All the materials can be found on BioNNE GitHub page[4] and on CodaLab[5] competition page.

3 Overview of Participation

3.1 Task 12b

This year, 26 teams participated in task 12b, submitting a total of 89 different systems across all three phases A, A+, and B. Specifically, 18, 8, and 16 teams competed in phases A, A+, and B, with 64, 34, and 54 distinct systems respectively. Eight of these teams were involved in all three phases. An overview of the technologies utilized by the teams is outlined in Table 4. Additional details for specific systems can be found in the workshop proceedings. As in previous years, the open-source system OAQA [51], which achieved top performance in older editions of BioASQ, was used as a baseline for phase B *exact answers*. This system is based on the UIMA framework and relies on traditional NLP and Machine Learning approaches and tools, such as MetaMap and LingPipe [5].

The MQU team from Macquarie University participated in all three phases of the task with five systems. Their systems relied on several open Large Language Models (LLMs) like llama-2, llama-3, gemini, and phi-3. Additionally, the team applied techniques like query expansion, re-ranking, and Retrieval Augmented Generation (RAG) on abstracts to improve their results. Another team participating in all phases is the team from the BSRC Alexander Fleming Institute. Their systems focused on sparse, dense and hybrid methods for document and snippet retrieval and LLMs with optimized prompts for exact and ideal answers. For document and query embedding as well as answer generation open LLMs were employed. Additionally, for Yes/No questions, a jury of complementary open LLMs perform majority voting to determine the final results.

The MiBi team from the Friedrich-Schiller-Universität Jena also participated in all phases with five systems. Their systems relied on LLM-based RAG. In phase A, their systems applied BM25 scoring to the article's title, abstract,

[4] https://github.com/nerel-ds/NEREL-BIO/tree/master/bio-nne/.
[5] https://codalab.lisn.upsaclay.fr/competitions/16464.

Table 4. Systems and approaches for task 12b. Systems for which no information was available at the time of writing are omitted.

Systems	Phase	Ref.	Approach
MQU	A, A+, B	[14]	llama-2, llama-3, gemini, phi-3, query expansion, re-ranking, RAG
BSRC	A, A+, B	[44]	open source LLMs, sparse/dense/hybrid retrieval
MiBi	A, A+, B	[46]	BM25, re-ranking, RAG, GPT-3.5, GPT-4, Mixtral-8x7B, DSPy LLM
UR	A, A+, B	[4]	Claude 3 Opus, GPT-3.5-turbo, Mixtral 8x7B, adapter fine-tuning, query expansion
UA	A, A+, B	[2]	BM25, PubMedBERT, BioLinkBERT, llama-2, llama-3, Nous-Hermes2-Mixtral, Gemma-2b,
CUHK-AIH	A, A+, B	[15]	BM25, llama-2, RAG
Gatech	A, A+, B	[55]	Mixtral, GPT-J, GPT-4, Llama2, resampling
Fudan-Atypon	A, A+, B	–	GPT3.5/4.0, ChatGLM, Spark, scenario prompt, LLama3-8B-instruct, query expansion
UNIPD	A	[21]	BM25, BiomedBERT, GPT-3.5, Gemini, NER
OPIX	A	–	BM25, GAT, re-ranking, DRUMS
HU	A	[56]	BM25, MedCPT, E5, GPT-3.5
NCU	B	[9]	GPT-4, RAG
UL	B	[3]	Mistral-7B-instruct, iterative fine-tuning
VCU	B	–	Synthia-13B, llama-3

and MeSH terms, while snippets were extracted either by GPT-3.5 chain-of-thought few-shot prompting or heuristically by re-ranking the title and chunks of up to three sentences from the abstract. They also performed re-ranking using pre-trained bi-encoders, cross-encoders, and lexical BM25 scoring. For phase A+, their systems were based on zero-shot prompting with GPT-3.5, GPT-4, or Mixtral-8x7B. The team also used the DSPy LLM programming framework for some runs. Furthermore, in phase B, their systems relied either on prompting an LLM (GPT-3.5 and GPT-4) with top-3 "ground truth" abstracts or all the "ground truth" snippets as context.

The UR team from the Universität Regensburg competed in all phases of the task with five systems. Their systems employed 1-shot and 10-shot learning combined with adapter fine-tuning for both commercial (Claude 3 Opus, GPT-3.5-turbo) and open source models (Mixtral 8x7B). In phase A, they relied on different 1-shot and 10-shot learning settings including plain, fine-tuned or models with additional context retrieved from Wikipedia, while in phases A+ and B, their systems relied solely on 10-shot learning.

The UA team from the Uni de Aveiro participated in all three phases of the task with five systems. In phase A, their systems followed a two-stage retrieval pipeline for the document retrieval using the traditional BM25 initially, followed by transformer-based neural re-ranking models, specifically PubMedBERT and BioLinkBERT. To enhance the BM25 results they used the BGE-M3 model, and reciprocal rank fusion to combine model outputs. For phases A+ and B, their systems employed instruction-based transformer models such as llama-2, llama-3, Nous-Hermes2-Mixtral, and a BioASQ fine-tuned version of Gemma 2B for conditioned zero-shot answer generation. Specifically, they utilized the top-5 most relevant articles to generate an ideal answer and used relevant snippets in Phase B.

The CUHK-AIH team from The Chinese University of Hong Kong also participated in all phases. Their systems are based on a RAG pipeline, with its base model being Llama2-chat-7B fine-tuned with LoRA using the training set. In Phase A, they first built indexes for all documents from PubMed Central using Pyserini to perform retrieval with BM25. In Phase A+, their systems further refined the retrieval by utilizing an ensemble retriever combining BM25 and vector similarity (bge-large-en embedding model). As for Phase B, the pipeline was similar to Phase A+, but the procedure was enhanced with the golden data.

The Gatech team from the Georgia Institute of Technology competed in all three phases. Their systems utilized a two-level information retrieval system and a QA system, based on pretrained LLMs including Mixtral, GPT-J, GPT-4 and llama-2, and prompt engineering. Specifically, their systems are based on LLM prompts with in-context few-shot examples to 1) parse the keywords in the given question to construct PubMed query and 2) solicit long and short answers for a question. Furthermore, they utilized post-processing techniques like resampling and malformed response detection to improve the performance.

The Fudan-Atypon team employed a two-stage IR model for phase A, similar to their previous work. In the first stage, they asked an LLM to extract words from the query, which were then used for query expansion. For Phase A+ and Phase B, they applied prompt engineering and four LLMs (GPT3.5/4.0, ChatGLM, Spark) for exact answer generation. For the final answers, they combined the distinct results to ensure stable performance. As for the ideal answer, they used a scenario prompt, which asked LLMs to reply in the way of a student majoring in biology. They fine-tuned a LLama3-8B-instruct to answer the questions in precise sentences, which has been proven to be one of the best methods.

In phase A, the UNIPD team from the University of Padova participated with five systems. As a first step, their systems utilized GPT-3.5 and Gemini to generate pseudo-documents that contain relevant information to a question and extract biomedical entities from these pseudo-documents using NER tools. Then, their systems follow a two-stage retrieval. In the first-stage, they used BM25 with the original queries concatenated together with the biomedical entities. In the second stage, they employed a BiomedBERT cross-encoder re-ranker, which was trained on the combination of golden standard data, synthetic data, as well as the LLM-generated pseudo-documents.

The OPIX team participated with two systems in phase A. Their systems followed a two-stage retrieval. They initially employed sparse document retrieval, followed by re-ranking which calculated the cosine similarity between the dense query and document representations and combined it with the cumulative scores of the sparse retrieval. Furthermore, their sparse retriever was based on BM25 and a graph attention neural network (GAT) that shared bidirectional information with a BERT model to enhance the re-ranking step. They utilised also the domain-specific UMLS knowledge graph, linking the entities mentioned on the PubMed documents to entity nodes of the UMLS graph. Their systems take pairs of queries and relevant KG subgraphs as input and bidirectionally fuses information from both modalities creating dense representations for the query and the node entities of the graph.

Also, the HU team competed in phase A with five systems. Their systems focused on using BM25 for sparse first-stage retrieval and combining a variety of dense, neural models via rank fusion for second-stage re-ranking. The neural re-rankers consist of i) two distinct cross-encoders (MedCPT and E5), one trained on a pairwise loss and one trained on both token-level and document-level features, as well as ii) one LLM approach (GPT-3.5), generating synthetic queries from documents returned by the first-stage retrieval and comparing their similarities to the original test query.

In phase B, the NCU team from the National Central University participated with five systems. Their systems utilized GPT-4 and RAG techniques to improve the retrieval process. They employed prompt engineering to refine the input queries guiding GPT-4 and improving its output accuracy and relevance. RAG was integrated to retrieve relevant biomedical documents, which were then incorporated into the generation process. The UL team from the Uni de Lisboa competed with one system in phase B. Their system focused on enhancing LLMs with external biomedical data. Specifically, they utilized the Mistral-7B-Instruct v0.1 model and an iterative process of fine-tuning using manually curated biomedical data alongside open-source resources. The VCU team from Virginia Commonwealth University participated with four different systems in phase B. Their systems are based on a zero-shot learning approach using generative LLMs, including Synthia-13B-GPTQ and llama-3. Their systems heavily relied on prompt engineering and answer processing.

3.2 Task Synergy 12

In the twelfth edition of BioASQ, four teams participated in the Synergy task (Synergy 12). These teams submitted results from 16 distinct systems. An overview of systems and approaches employed is provided in Table 5.

Table 5. Systems and their approaches for task Synergy. Systems for which no description was available at the time of writing are omitted.

System	Ref.	Approach
BSRC	[44]	open source LLMs, sparse/dense/hybrid retrieval
UR	[4]	GPT-3.5-turbo, GPT-4, query expansion
Gatech	[55]	Mixtral, GPT-J, GPT-4, Llama2, resampling

In particular, the BSRC Alexander Fleming team participated with four systems. Similar to task b, their systems focused on LLMs with optimized prompts and majority voting. Also, the UR team from the Universität Regensburg competed with two systems. Their systems employed 2-shot and zero-shot learning with GPT-3.5-turbo and GPT-4. Furthermore, their systems utilized handcrafted examples for the 2-shot learning as well as incorporated query expansion methods. The Gatech team from the Georgia Institute of Technology participated in five systems. As with task b, their systems relied on pre-trained LLMs and prompt engineering. More detailed descriptions for some of the systems are available at the proceedings of the workshop.

3.3 Task MultiCardioNER

31 teams registered for the MultiCardioNER task, out of which 7 teams submitted at least one run of their predictions. Specifically, 6 teams participated in the CardioDis subtrack, while 5 teams participated in the MultiDrug subtrack (with one of those teams participating only in the Spanish part). Overall, a total of 70 runs were submitted, with each team being allowed up to 5 runs per subtrack and language.

Table 6 gives an overview of the methodologies used by the participants in each of the sub-tasks. Following the trend of previous similar shared tasks [28–30], all participants used some variant of Transformers-based models, with RoBERTa [32] models being the most popular. Other than that, ensembles were quite popular and provided good results (e.g. BIT.UA [23]), as were the use of custom datasets and dictionaries (e.g. Enigma [1]), data augmentation or window sliding (e.g. Data Science TUW [49]). It is also noteworthy the way in which the teams incorporated the cardiology-specific data, with some teams trying to mesh it into their training data in different ways (e.g. PICUSLab [47]) and others using only the mixed-specialty training data (e.g. NOVALINCS [17]). Finally, an interesting aspect of the MultiDrug subtrack is that, while the most common approach was to focus creating separate, language-specific models, there were some teams who tried to create purely multilingual models attempting to optimize the performance for all three languages at once, such as the ICUE team [26].

3.4 Task BioNNE

26 teams registered for the BioNNE task in CodaLab, out of which 5 teams submitted at least one run of their predictions. Overall, a total of 155 runs were submitted. An overview of the approaches is provided in Table 7. Two systems for which no information was available at the time of writing are omitted.

Team **fulstock** employed the BINDER model [53], which uses XLM-RoBERTa [10] as its backbone. The team experimented with various entity type descriptions (prompts) for BINDER learning. These prompts included: keyword (name of the entity type), 2, 5, or 10 most frequent component words for entity

Table 6. General overview of the approaches presented by participants for the MultiCardioNER task. "*TEMIST corpora" refers to the joint version of the DisTEMIST, SympTEMIST, ProcTEMIST and DrugTEMIST corpora.

Team	Ref.	Task	Approaches
BIT.UA	[23]	CardioDis	Ensemble of RoBERTa models with multi-head CRF and differences in the data used for training (only DisTEMIST or DisTEMIST + CardioCCC)
Data Science TUW	[49]	CardioDis	Transformer-based models with different pretraining settings, data augmentation and window sliding
		MultiDrug	Multilingual and language-specific Transformers with different pretraining settings, data augmentation and window sliding
Enigma	[1]	CardioDis	CLIN-X-ES model fine-tuned on the entire task data + custom clinical dataset
		MultiDrug	Multilingual and language-specific Transformers fine-tuned on the entire task data + custom drug dictionary
ICUE	[26]	MultiDrug	Multilingual and language-specific BERT models with re-training, post-processing rules + GPT 3.5
NOVALINCS	[17]	CardioDis	RoBERTa model fine-tuned on the standalone DisTEMIST corpus vs. joint *TEMIST corpora
		MultiDrug	RoBERTa model fine-tuned on the standalone DrugTEMIST corpus vs. joint *TEMIST corpora
PICUSLab	[47]	CardioDis	Ensemble of Transformer-based models trained on different datasets, including an augmented version of CardioCCC + post-processing via string matching
Siemens	[11]	CardioDis	Fine-tuned general domain BERT model
		MultiDrug	Fine-tuned language-specific general domain BERT models

Table 7. Overview of the approaches presented by participants for the BioNNE task. EN stands for the English-oriented and RU for the Russian-oriented tracks.

Team	Ref.	Track	Approaches
fulstock	–	Bilingual, EN, RU	BINDER, XLM-RoBERTa
wenxinzh	[54]	EN	Mixtral, spaCy NER, UMLS
hasin.rehana	[45]	Bilingual, EN, RU	PubMedBERT, SBERT-Large-NLU-RU, BERT-Base-Multilingual-uncased, UMLS

type in the training data, contextual prompt (a sentence example with the target entity), and lexical prompt (a sentence example where the target entity is masked with the entity label) [48]. The model was trained over 64 epochs.

Team **wenxinzh** [54] combined a pretrained Mixtral model [22] with a spaCy NER model trained on the BC5CDR corpus [27]. They also adapted and customized rules based on UMLS (Unified Medical Language System) semantic types. The system first utilizes Mixtral and en_ner_bc5cdr_md to extract potential entities for each category from the text and then determines their final entity types by finding the associated UMLS semantic types.

Team **hasin.rehana** [45] implemented the BIO-tagging scheme, applying six levels of BIO-tagging. They added six classification layers to the base model, each dedicated to outputting a specific level of NER tags. The original dataset's eight classes were expanded to 17 to accommodate the BIO-tagging scheme. For vocabulary expansion, they utilized the UMLS Metathesaurus to extract relevant and related concepts. For English NNER, they used the pre-trained PubMedBERT for contextualized word embeddings [18]; for Russian NNER, they employed a pre-trained SBERT-Large-NLU-RU model; and for Bilingual NNER, they utilized BERT-Base-Multilingual-uncased [13].

4 Results

4.1 Task 12b

This section presents the evaluation measures and preliminary results for the task 12b. These results are preliminary, as the final results will be available after the manual assessment of the system responses by the BioASQ team of experts and the enrichment of the ground truth with potential additional relevant items, answer elements, and/or synonyms, which is still in progress.

Phase A: The Mean Average Precision (MAP) was used for evaluation on document retrieval. In particular, since BioASQ8, MAP calculation is based on a modified version of Average Precision (AP) that considers both the limit of 10 elements allowed per question in each submission and the actual number of golden elements that is often less than 10 in practice [40]. For snippets, where a single ground-truth snippet may overlap with several submitted ones, the interpretation of MAP is less straightforward. Hence, since BioASQ9, we use the F-measure which is based on character overlaps[6] [38]. Table 8 presents some indicative results for document retrieval in batch 1. The full 12b results for phase A are available online[7].

Phases A+ and B: The official ranking for systems providing *ideal answers* is based on manual scores assigned by the BioASQ team of experts that assesses each *ideal answer* in the responses [6]. The final position of systems providing *exact answers* is based on their average ranking in the three question types where *exact answers* are required, that is "yes/no", "list", and "factoid". Summary questions for which no *exact answers* are submitted are not considered in this ranking. In particular, the mean F1 measure is used for the ranking in list questions, the Mean Reciprocal Rank (MRR) is used for the ranking in factoid questions, and the F1 measure, macro-averaged over the classes of yes and no, is used for yes/no questions. Tables 9 and 10 present some indicative results on *exact answer* extraction. The full 12b results for both phase A+[8] and B[9] are available online.

[6] http://participants-area.bioasq.org/Tasks/b/eval_meas_2022/.
[7] http://participants-area.bioasq.org/results/12b/phaseA/.
[8] http://participants-area.bioasq.org/results/12b/phaseAplus/.
[9] http://participants-area.bioasq.org/results/12b/phaseB/.

Table 8. Preliminary results for document retrieval in batch 1 of phase A of task 12b. Only the top-6 systems are presented, based on MAP.

System	Mean Precision	Mean Recall	Mean F-measure	MAP	GMAP
bioinfo-4	0.1039	0.3124	0.1485	**0.2067**	0.0016
bioinfo-3	0.1009	0.3047	0.1444	0.2024	0.0013
bioinfo-1	0.1156	0.3171	0.1581	0.2018	0.0015
bioinfo-2	**0.1294**	**0.3369**	**0.1728**	0.2006	**0.0019**
bioinfo-0	0.1151	0.3055	0.1570	0.1800	0.0009
dmiip2024_3	0.0706	0.2514	0.1039	0.1612	0.0007

Table 9. Results for batch 1 for *exact answers* in phase A+ of task 12b. Only the top-6 systems based on Yes/No F1 are presented.

System	Yes/No		Factoid			List		
	F1	Acc.	Str. Acc.	Len. Acc.	MRR	Prec.	Rec.	F1
UR-IW-3	**0.9167**	**0.920**	0.0952	0.0952	0.0952	0.4016	0.4778	0.4089
Gatech comp...	0.8397	0.840	0.1429	0.1429	0.1429	0.4452	0.3415	0.3661
GTBioASQsys3	0.8397	0.840	0.1429	0.1429	0.1429	0.2421	0.1765	0.1866
UR-IW-4	0.8397	0.840	0.0476	0.0952	0.0714	0.3948	0.4063	0.3798
UR-IW-2	0.8397	0.840	0.0952	0.0952	0.0952	**0.5250**	**0.4914**	**0.4808**
UR-IW-5	0.7987	0.800	0.0952	0.0952	0.0952	0.4119	0.4182	0.3976

Table 10. Results for batch 3 for *exact answers* in phase B of task 12b. Only the top-6 systems based on Yes/No F1 and the BioASQ Baseline are presented.

System	Yes/No		Factoid			List		
	F1	Acc.	Str. Acc.	Len. Acc.	MRR	Prec.	Rec.	F1
mibi_rag_snippet	**1.00**	**1.00**	0.2308	0.2308	0.2308	0.4984	0.5157	0.5052
RMC_append_sn...	**1.00**	**1.00**	0.3077	0.3077	0.3077	0.4158	0.4475	0.3955
Fleming-3	**1.00**	**1.00**	0.2308	0.2692	0.2404	0.5424	0.5532	0.5413
IISR 4th submit	**1.00**	**1.00**	0.4231	0.4231	0.4231	0.5452	0.5187	0.5247
dmiip2024_2	**1.00**	**1.00**	0.1923	0.4231	0.2949	0.5114	0.5055	0.4715
GTBioASQsys2	0.9577	0.9583	0.3846	0.3846	0.3846	0.5107	0.4774	0.4763
BioASQ_Baseline	0.4338	0.4583	0.0769	0.1538	0.1090	0.1999	0.2938	0.2094

The top performance of the participating systems in *exact answer* generation for each type of question during the twelve years of BioASQ is presented in Fig. 1. The preliminary results for task 12b, reveal that the participating systems keep achieving high scores in answering all types of questions, despite the addition of two new experts to the BioASQ team. In batch 3, for instance, presented in Table 10, several systems manage to correctly answer literally all yes/no questions. High performance, beyond 0.95 in macro F1 is also observed for yes/no questions in the remaining batches. More consistent performance is also observed

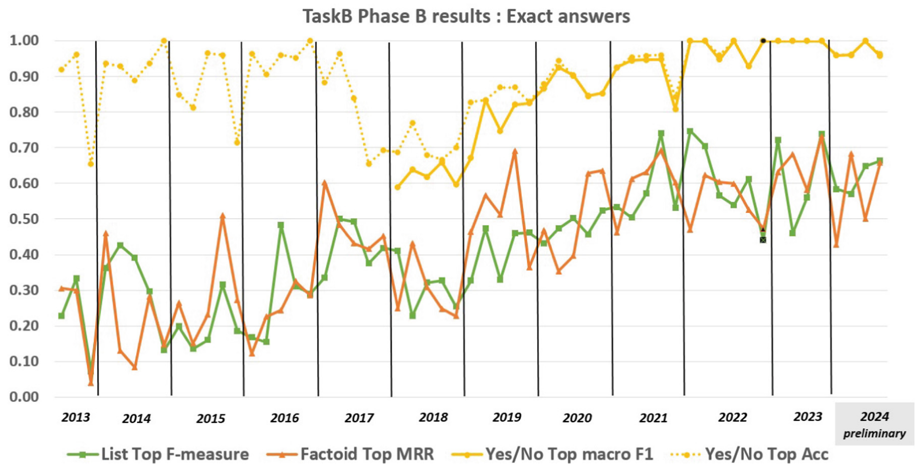


Fig. 1. The evaluation scores of the best-performing systems in task B, Phase B, for *exact answers*, across the twelve years of BioASQ. Since BioASQ6, accuracy (Acc) was replaced by macro F1 as the official measure for Yes/No questions. The black dots indicate an additional batch with questions from new experts [38].

in the preliminary results for list questions compared to the previous years, but there is still more room for improvement, as done for factoid questions where the performance across the batches fluctuates more.

4.2 Task Synergy 12

In task Synergy 12, no relevant material was initially available for new questions. For old questions, however, feedback from previous rounds was provided per question, that is the documents and snippets submitted by the participants with manual annotations of their relevance. Hence, the documents and snippets of the feedback, that have already been assessed and released, were not considered valid for submission in the subsequent rounds. As in task 12b, the evaluation measures for document and snippet retrieval are MAP and F-measure respectively.

In addition, due to the developing nature of the topic, no answer is available for all of the open questions in each round. Therefore only the questions indicated as “answer ready” were evaluated for *exact* and *ideal answers* in each round. Regarding the *ideal answers*, the systems were ranked according to manual scores assigned to them by the BioASQ experts during the assessment of systems responses as in phase B of task B [6]. As regards evaluation for the *exact answers*, similarly to task 12b, the mean F1 measure, the Mean Reciprocal Rank (MRR), and the macro F1 measure are used for the ranking in list, factoid, and yes/no questions respectively. Any *exact* or *ideal answer* that was assessed as ground-truth quality by the experts, was included in the feedback and provided to the participants before the next round.

Table 11. Results for document retrieval of the first round of the Synergy 12 task.

System	Mean precision	Mean Recall	Mean F-Measure	MAP	GMAP
dmiip3	**0.4043**	**0.4718**	**0.3558**	**0.4636**	0.1783
dmiip1	0.3899	0.4697	0.3472	0.4535	0.1697
dmiip4	0.3971	0.4674	0.3508	0.4493	**0.1928**
dmiip5	0.3971	0.4674	0.3508	0.4493	**0.1928**
dmiip2	0.3826	0.4690	0.3416	0.4427	0.1814
Fleming-4	0.3176	0.3385	0.2544	0.2342	0.0324
Fleming-2	0.2664	0.2524	0.2235	0.2152	0.0056
Fleming-1	0.2756	0.2476	0.2246	0.2110	0.0056
Fleming-3	0.2695	0.2696	0.2163	0.1985	0.0157

Some indicative results for the Synergy task are presented in Table 11. The full 12b results are available online[10]. Overall, the collaboration between participating biomedical experts and question-answering systems allowed the progressive identification of relevant material and extraction of *exact* and *ideal answers* for several open questions for developing problems, such as Colorectal Cancer, pediatric sepsis, Duchenne Muscular Dystrophy, and COVID-19. In total, after the four rounds of Synergy 12, enough relevant material was identified to provide an answer to about 78% of the questions. In addition, about 51% of the questions had at least one *ideal answer*, submitted by the systems, which was considered of ground-truth quality by the respective expert.

4.3 Task MultiCardioNER

All in all, the top scores for each subtrack were:

– **CardioDis**. The team BIT.UA attained the top position with an ensemble of RoBERTa-based models (roberta-es-clinical-trials-ner) that also uses multi-head CRF. Their runs integrated the provided datasets in different ways, with the highest scores being achieved by the models that use both the DisTEMIST and CardioCCC data. Their best run achieved an F1-score of 0.8199 and a recall of 0.8243. The team with the next best F1-score (0.8049) is Enigma, who uses a CLIN-X-ES model also fine-tuned on the DisTEMIST and CardioCCC data. Interestingly, the team PICUSLab achieves the best precision (0.8886) by a wide margin combining the predictions of multiple models trained on different parts of the data (including an augmented version of the CardioCCC corpus) and then using string matching techniques to enhance the final predictions.

[10] http://participants-area.bioasq.org/results/synergy_v2024/.

– **MultiDrug**. In Spanish, the best F1-score is achieved by the ICUE team (0.9277), who also achieved the best recall (0.9412). Meanwhile, in English and Italian the winner team is Enigma, with an F1-score of 0.9223 and 0.8842, respectively.

The results for the CardioDis subtrack are shown in Table 12, while the results for the MultiDrug subtrack are presented in Table 13 for Spanish, Table 14 for English and Table 15 for Italian. Due to space limitations, only the top-6 systems are presented, with the complete results being available in the MultiCardioNER overview paper [31].

In conclusion, the task's results are quite good and varied, with scores ranging from 0.9277 (by the ICUE team in the Spanish MultiDrug subtrack) to 0.2201 (by the DataScienceTUW team, who had some problems with the submission, also in the Spanish MultiDrug subtrack). Overall, the results for the MultiDrug subtrack are higher than those for the CardioDis subtrack, which was to be expected since drugs, as an entity type, are simpler and more straightforward than diseases.

As stated earlier, in terms of methodology, there's a definite trend of using pre-trained Transformer-based systems (with a preference for RoBERTa models, perhaps due to their availability in Spanish), with most participants going beyond mere finetuning. Many of the presented runs incorporate new layers over their initial system results, be it an ensemble of multiple models and their predictions, multi-head CRFs, window sliding or using some kind of post-processing.

However, what seems to really make a difference in MultiCardioNER is the use of the cardiology-specific data (the CardioCCC dataset), which is one of the shared task's main research points. All top-performing systems incorporate the released 258 documents from the CardioCCC corpus in some way. Meanwhile, participants that only use the DisTEMIST and DrugTEMIST corpora (which are made up of clinical case reports from varied clinical specialties) are able to achieve a really high precision but a much lower recall, thus achieving a not-so-high F1-score. This seems to indicate that, while these systems are able to retrieve many clinical entities correctly (i.e. high precision), they fail to recover those entities that are specific to the cardiology domain (i.e. low recall).

Furthermore, comparing the results of the DisTEMIST shared task [35], which also focused on diseases, with the CardioDis subtrack, shows an improvement in the overall results in this new task. All of this seems to point towards the importance of using data belonging to the clinical specialty that we plan to apply our systems to, even within domains that are already quite specific as is the clinical domain. Still, it is true that, compared with DisTEMIST, this task offers a higher volume of data. While there seems to be a positive correlation with the use of domain-specific data, whether these improvements can actually be attributed to the domain adaptation aspect or to simply having more data remains to be seen and is a question for further research.

As for the multilingual aspect of the track, the results for all three languages are quite comparable, with Italian being somewhat below Spanish and English. This difference might be explained by the fact that, while there are many pre-

Table 12. Results of the MultiCardioNER CardioDis subtrack. Only the top-6 systems are presented. The best result is bolded, and the second-best is underlined.

Team Name	Run name	Precision	Recall	F1
BIT.UA	run1-all-full	0.8155	**0.8243**	**0.8199**
BIT.UA	run0-top5-full	0.811	<u>0.8181</u>	<u>0.8145</u>
Enigma	3-system-CLIN-X-ES-pretrained	0.8016	0.8082	0.8049
Enigma	2-system-CLIN-X-ES-14	0.8052	0.8007	0.803
PICUSLab	aug_fus_sub2	0.7794	0.803	0.791
BIT.UA	run4-all	0.7981	0.7827	0.7903

Table 13. Results of the MultiCardioNER MultiDrug subtrack in Spanish. Only the top-6 systems are presented. The best result is bolded, and the second-best is underlined.

Team Name	Run name	Precision	Recall	F1
ICUE	run2_single_pp	0.9146	**0.9412**	**0.9277**
ICUE	run4_GPT_translation	0.9146	0.9412	0.9277
ICUE	run5_GPT_translation_all	0.9146	0.9412	0.9277
Enigma	3-system-SpanishRoBERTa	0.913	<u>0.9348</u>	<u>0.9238</u>
Enigma	1-system-XLMR	0.904	0.9208	0.9123
Enigma	2-system-XLMR-filtering	<u>0.9148</u>	0.9005	0.9076

Table 14. Results of the MultiCardioNER MultiDrug subtrack in English. Only the top-6 systems are presented. The best result is bolded, and the second-best is underlined.

Team Name	Run name	Precision	Recall	F1
Enigma	3-system-BioLinkBERT	0.8981	**0.9477**	**0.9223**
ICUE	run2_single_pp	**0.9086**	0.9128	<u>0.9107</u>
ICUE	run4_GPT_translation	0.9086	0.9128	0.9107
Enigma	1-system-XLMR	0.8823	0.9233	0.9023
Enigma	2-system-XLMR-filtering	<u>0.9031</u>	0.8989	0.901
Enigma	5-system-XLMR-filtering-dict2	0.8698	0.9047	0.8869

trained models available in Spanish and English that were used by participants, this is not the case for Italian. In fact, the only Italian-specific models used were a version of BERT in Italian (that is, a general domain model) and BioBIT [8], a model that is specific to the biomedical domain trained on machine-translated PubMed abstracts.

In contrast, participants were able to use a wider variety of models for English, such as BioLinkBERT [52] or SciBERT [7], and Spanish. An approach to solve this lack of clinical models in Italian followed by some participants was to further pre-train existing Spanish models using Italian and multilingual data, which made the Enigma team achieve the three top-scoring runs in the Ital-

Table 15. Results of the MultiCardioNER MultiDrug subtrack in Italian. Only the top-6 systems are presented. The best result is bolded, and the second-best is underlined.

Team Name	Run name	Precision	Recall	F1
Enigma	1-system-XLMR	0.884	0.8844	**0.8842**
Enigma	3-system-Italian-Spanish-RoBERTa	0.8723	<u>0.8956</u>	<u>0.8838</u>
Enigma	2-system-XLMR-filtering	<u>0.9016</u>	0.8606	0.8806
Siemens	run1_IMR	0.8891	0.8689	0.8789
ICUE	run4_GPT_translation	**0.9114**	0.8461	0.8776
ICUE	run5_GPT_translation_all	0.9114	0.8461	0.8776

ian track. Multilingual models such as mDeBERTa [19,20] were also used by participants.

4.4 Task BioNNE

The primary evaluation metric utilized in this study is the F_1-score, which is computed using the following formula: $F_1 = \frac{1}{n}\sum_{c\in C} F1_{rel_c}$, where C represents the set of classes {`FINDING`, `DISO`, `INJURY_POISONING`, `PHYS`, `DEVICE`, `LABPROC`, `ANATOMY`, `CHEM`}, n is the size of C, and $F1_{rel_c}$ denotes the macro F_1-score averaged across all relevance classes.

Table 16. Results (F_1 scores on the test sets) of bilingual and monolingual subtasks. The best result in each task is bolded.

Model	Both (Track 1)	English (Track 2)	Russian (Track 3)
fulstock	**0.7044**	**0.6181**	**0.6981**
hasin.rehana	0.5053	0.5636	0.6007
wenxinzh	–	0.348	–

We summarized the performance of the above-mentioned teams in Table 16. The fulstock team with the fine-tuned BINDER model achieved the highest F_1 scores across all tracks, with 0.704 for the bilingual track, 0.618 for the English-oriented track, and 0.698 for the Russian-oriented track. In contrast, a pre-trained LLM, specifically the Mixtral model combined with the NER model for flat entities, achieved an F_1 score of 0.34797 for the English-oriented track. This score can be considered indicative of zero-shot evaluation, highlighting the limitations due to the absence of supervised training and the inadequacy of biomedical-specific training data in LLMs such as Mixtral. More results, along with baselines, are available in the BioNNE overview paper [12].

5 Conclusions

This paper provides an overview of the twelfth BioASQ challenge. This year, BioASQ consisted of four tasks: (1) Task 12b on biomedical semantic question answering in English and (2) Synergy 12 on question answering for developing problems, both already established from previous BioASQ versions, (3) the new task MultiCardioNER on the automatic detection of disease and drug mentions on cardiology clinical case reports in Spanish, Italian, and English, and (4) the new task BIONNE on biomedical nested NER in English and Russian.

The preliminary results for task 12b reveal the high performance of the top participating systems, predominantly for yes/no answer generation, despite the extension of the expert team with two new experts. However, room for improvement is still available, particularly for factoid and list questions, where the performance is less consistent. The results of the new Phase A+ also reveal that state-of-the-art QA approaches can achieve high performance, even without access to manually selected relevant material. Still, providing such material leads to answers of improved quality. This edition of the Synergy task as well, revealed that state-of-the-art systems, despite still having room for improvement, can be a useful tool for biomedical experts who need specialized information for addressing open questions in the context of several developing problems.

The new task MultiCardioNER presented two new challenging subtasks about annotations clinical case reports in Spanish, English, Italian with disease and drug mentions. Building on the work laid out in previous shared tasks like DisTEMIST [35], this task introduces the nuance of creating clinical Named Entity Recognition systems specifically for the cardiology domain. In addition, it expands the range of the task beyond Spanish by introducing a subtrack that also involves English and Italian text. In order to do this, two new datasets are released: the DrugTEMIST corpus, that includes drug mentions in Spanish, English and Italian in a group of clinical case reports of varied medical specialties, and the CardioCCC corpus, a collection of cardiology clinical case reports with disease and drug annotations. The results highlight the importance of having data specific to the language and specialty the systems are going to be applied in, even within domains that are already quite specific like the clinical one.

The ever-increasing focus of participating systems on deep neural approaches and Large Language Models, already apparent in previous editions of the challenge, is also observed this year. Most of the proposed approaches built on state-of-the-art neural architectures (BERT, PubMedBERT, BioBERT, BART etc.) adapted to the biomedical domain and specifically to the tasks of BioASQ. This year, in particular, several teams investigated approaches based on Generative Pre-trained Transformer (GPT) models and Retrieval Augmented Generation (RAG) for the BioASQ tasks.

The BioNNE task centered on extracting the eight most common biomedical entities in Russian and English from PubMed abstracts while accommodating potential nested structures. The top-performing approach employed a bi-encoder framework that leverages contrastive learning to map text spans and entity types into a common vector representation space. The performance of pre-trained LLMs without fine-tuning exhibited significantly lower results, underscoring the necessity for specialized training data.

Overall, several systems managed competitive performance on the challenging tasks offered in BioASQ, as in previous versions of the challenge, and the top performing of them were able to improve over the state-of-the-art performance from previous years. BioASQ keeps pushing the research frontier in biomedical semantic indexing and question answering for eleven years now, offering both well-established and new tasks. Aligned with the direction of extending beyond the English language and biomedical literature, which started with the task MESINESP [16] and continued consistently ever since, this year BioASQ was further extended with two new tasks, MultiCardioNER [31] and BioNNE [12]. In addition, this year we introduced a new phase in the QA task 12b (phase A+) allowing the assessment of systems that produce answers directly, without access to manually selected relevant material. The future plans for the challenge include a further extension of the benchmark data for question answering through a community-driven process, extending the community of biomedical experts involved in the Synergy task, as well as extending the resources considered in the BioASQ tasks, both in terms of documents types, languages, and more focused sub-domains of biomedicine.

Acknowledgments. Google was a proud sponsor of the BioASQ Challenge in 2023. Ovid is also sponsoring this edition of BioASQ. The twelfth edition of BioASQ is also sponsored by Elsevier. Atypon Systems Inc. is also sponsoring this edition of BioASQ. The MEDLINE/PubMed data resources considered in this work were accessed courtesy of the U.S. National Library of Medicine. BioASQ is grateful to the CMU team for providing the *exact answer* baselines for task 12b. The MultiCardioNER track was funded by Spanish and European projects such as DataTools4Heart (Grant Agreement No. 101057849), AI4HF (Grant Agreement No. 101080430), BARITONE (Proyectos de Transición Ecológica y Transición Digital 2021. Expediente Nº TED2021-129974B-C21) and AI4ProfHealth (PID2020-119266RA-I00). The work on the BioNNE task was supported by the Russian Science Foundation [grant number 23-11-00358].

References

1. Aksenova, A., Datseris, A., Vassileva, S., Boytcheva, S.: Transformer-based disease and drug named entity recognition in multilingual clinical texts: MultiCardioNER challenge. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) CLEF Working Notes (2024)
2. Almeida, T., Jonker, R., Reis, J., Almeida, J., Matos, S.: From retrieval to answer generation: insights from BioASQ 12 task B. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) CLEF Working Notes (2024)
3. Anaya, C., Fernandes, M., Couto, F.: LLM fine-tuning with biomedical open-source data. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) CLEF Working Notes (2024)

4. Ateia, S., Kruschwitz, U.: Can open-source LLMs compete with commercial models? Exploring the few-shot performance of current GPT models in biomedical tasks. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) CLEF Working Notes (2024)
5. Baldwin, B., Carpenter, B.: Lingpipe. Available from World Wide Web (2003). http://alias-i.com/lingpipe
6. Balikas, G., et al.: Evaluation framework specifications. Project deliverable D4.1, UPMC (2013)
7. Beltagy, I., Lo, K., Cohan, A.: SciBERT: pretrained language model for scientific text. In: EMNLP (2019)
8. Buonocore, T.M., Crema, C., Redolfi, A., Bellazzi, R., Parimbelli, E.: Localizing in-domain adaptation of transformer-based biomedical language models. J. Biomed. Inform. **144**, 104431 (2023)
9. Chih, B.C., Han, J.C., Tzong-Han Tsai, R.: NCU-IISR: enhancing biomedical question answering with GPT-4 and retrieval augmented generation in BioASQ 12b phase B. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) CLEF Working Notes (2024)
10. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale. CoRR abs/1911.02116 (2019). http://arxiv.org/abs/1911.02116
11. Danu, M.D., Marica, V.G., Suciu, C., Itu, L.M., Farri, O.: Multilingual clinical NER for diseases and medications recognition in cardiology texts using BERT embeddings. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) CLEF Working Notes (2024)
12. Davydova, V., Loukachevitch, N., Tutubalina, E.: Overview of BioNNE task on biomedical nested named entity recognition at BioASQ 2024. In: CLEF Working Notes (2024)
13. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805 (2018). http://arxiv.org/abs/1810.04805
14. Galat, D., Moshkin, S.: Refining zero-short approaches for biomedical question answering. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) CLEF Working Notes (2024)
15. Gao, Y., Zong, L., Li, Y.: Enhancing biomedical question answering with parameter-efficient fine-tuning and hierarchical retrieval augmented generation. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) CLEF Working Notes (2024)
16. Gasco, L., et al.: Overview of BioASQ 2021-MESINESP track. Evaluation of advance hierarchical classification techniques for scientific literature, patents and clinical trials (2021)
17. Gonçalves, R., Lamúrias, A.: Team NOVA LINCS @ BIOASQ12 MultiCardioNER track: entity recognition with additional entity types. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) CLEF Working Notes (2024)
18. Gu, Y., et al.: Domain-specific language model pretraining for biomedical natural language processing. ACM Trans. Comput. Healthc. (HEALTH) **3**(1), 1–23 (2021)
19. He, P., Gao, J., Chen, W.: DeBERTaV3: improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing (2021)
20. He, P., Liu, X., Gao, J., Chen, W.: DeBERTa: decoding-enhanced BERT with disentangled attention. In: International Conference on Learning Representations (2021). https://openreview.net/forum?id=XPZIaotutsD

21. Huang, B.W.: Generative large language models augmented hybrid retrieval system for biomedical question answering. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) CLEF Working Notes (2024)
22. Jiang, A.Q., et al.: Mixtral of experts (2024)
23. Jonker, R., Almeida, T., Matos, S.: BIT.UA at MultiCardioNER: adapting a multi-head CRF for cardiology. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) CLEF Working Notes (2024)
24. Krithara, A., Nentidis, A., Bougiatiotis, K., Paliouras, G.: BioASQ-QA: a manually curated corpus for biomedical question answering. Sci. Data **10**(1), 170 (2023)
25. Krithara, A., Nentidis, A., Paliouras, G., Krallinger, M., Miranda, A.: BioASQ at CLEF2021: large-scale biomedical semantic indexing and question answering. In: Hiemstra, D., Moens, M.-F., Mothe, J., Perego, R., Potthast, M., Sebastiani, F. (eds.) ECIR 2021. LNCS, vol. 12657, pp. 624–630. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-72240-1_73
26. Lee, C., Simpson, T.I., Posma, J.M., Lain, A.D.: Comparative analyses of multilingual drug entity recognition systems for clinical case reports in cardiology. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) CLEF Working Notes (2024)
27. Li, J., et al.: BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database J. Biol. Databases Curation **2016** (2016). https://doi.org/10.1093/database/baw068
28. Lima-López, S., Farré-Maduell, E., Brivá-Escalada, V., Gascó, L., Krallinger, M.: MEDDOPLACE Shared Task overview: recognition, normalization and classification of locations and patient movement in clinical texts. Procesamiento del Lenguaje Natural **71**, 301–311 (2023)
29. Lima-López, S., Farré-Maduell, E., Gasco-Sánchez, L., Rodríguez-Miret, J., Krallinger, M.: Overview of SympTEMIST at BioCreative VIII: corpus, guidelines and evaluation of systems for the detection and normalization of symptoms, signs and findings from text. In: Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models (2023)
30. Lima-López, S., et al.: Overview of MedProcNER task on medical procedure detection and entity linking at BioASQ 2023. In: Working Notes of CLEF 2023 (2023)
31. Lima-López, S., et al.: Overview of MultiCardioNER task at BioASQ 2024 on medical speciality and language adaptation of clinical NER systems for Spanish, English and Italian. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) CLEF Working Notes (2024)
32. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
33. Loukachevitch, N., et al.: NEREL: a Russian information extraction dataset with rich annotation for nested entities, relations, and wikidata entity links. Lang. Resour. Eval. **58**, 547–583 (2023)
34. Loukachevitch, N., et al.: NEREL-BIO: a dataset of biomedical abstracts annotated with nested named entities. Bioinformatics **39**(4), btad161 (2023). https://doi.org/10.1093/bioinformatics/btad161
35. Miranda-Escalada, A., et al.: Overview of DisTEMIST at BioASQ: automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources (2022)
36. Nentidis, A., et al.: Overview of BioASQ 2023: the eleventh BioASQ challenge on large-scale biomedical semantic indexing and question answering. In: Arampatzis, A., et al. (eds.) CLEF 2023. LNCS, vol. 14163, pp. 227–250. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-42448-9_19

37. Nentidis, A., Katsimpras, G., Krithara, A., Paliouras, G.: Overview of BioASQ tasks 12b and Synergy12 in CLEF2024. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) CLEF Working Notes (2024)
38. Nentidis, A., et al.: Overview of BioASQ 2021: the ninth BioASQ challenge on large-scale biomedical semantic indexing and question answering. In: Candan, K.S., et al. (eds.) CLEF 2021. LNCS, vol. 12880, pp. 239–263. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-85251-1_18
39. Nentidis, A., et al.: Overview of BioASQ 2022: the tenth BioASQ challenge on large-scale biomedical semantic indexing and question answering. In: Barrón-Cedeño, A., et al. (eds.) CLEF 2022. LNCS, vol. 13390, pp. 337–361. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-13643-6_22
40. Nentidis, A., et al.: Overview of BioASQ 2020: the eighth BioASQ challenge on large-scale biomedical semantic indexing and question answering. In: Arampatzis, A., et al. (eds.) CLEF 2020. LNCS, vol. 12260, pp. 194–214. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58219-7_16
41. Nentidis, A., Krithara, A., Paliouras, G., Farre-Maduell, E., Lima-Lopez, S., Krallinger, M.: BioASQ at CLEF2023: the eleventh edition of the large-scale biomedical semantic indexing and question answering challenge. In: Kamps, J., et al. (eds.) ECIR 2023, Part III. LNCS, vol. 13982, pp. 577–584. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-28241-6_66
42. Nentidis, A., Krithara, A., Paliouras, G., Gasco, L., Krallinger, M.: BioASQ at CLEF2022: the tenth edition of the large-scale biomedical semantic indexing and question answering challenge. In: Hagen, M., et al. (eds.) ECIR 2022. LNCS, vol. 13186, pp. 429–435. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-99739-7_53
43. Nentidis, A., et al.: BioASQ at CLEF2024: the twelfth edition of the large-scale biomedical semantic indexing and question answering challenge. In: Goharian, N., et al. (eds.) ECIR 2024. LNCS, vol. 14612, pp. 490–497. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-56069-9_67
44. Panou, D., Dimopoulos, A., Reczko, M.: Farming open LLMs for biomedical question answering. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) CLEF Working Notes (2024)
45. Rehana, H., et al.: Nested named entity recognition using multilayer BERT-based model. In: CLEF Working Notes (2024)
46. Reimer, J.H., Bondarenko, A., Hagen, M., Viehweger, A.: MiBi at BioASQ 2024: retrieval-augmented generation for answering biomedical questions. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) CLEF Working Notes (2024)
47. Romano, A., Riccio, G., Postiglione, M., Moscato, V.: Identifying cardiological disorders in spanish via data augmentation and fine-tuned language models. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) CLEF Working Notes (2024)
48. Rozhkov, I., Loukachevitch, N.: Prompts in few-shot named entity recognition. Pattern Recogn. Image Anal. **33**(2), 122–131 (2023)
49. Styll, P., Campillos-Llanos, L., Kusa, W., Hanbury, A.: Cross-linguistic disease and drug detection in cardiology clinical texts: methods and outcomes. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) CLEF Working Notes (2024)
50. Tsatsaronis, G., et al.: An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinform. **16**, 138 (2015)

51. Yang, Z., Zhou, Y., Eric, N.: Learning to answer biomedical questions: OAQA at bioASQ 4B. In: ACL 2016, p. 23 (2016)
52. Yasunaga, M., Leskovec, J., Liang, P.: LinkBERT: pretraining language models with document links. In: Association for Computational Linguistics (ACL) (2022)
53. Zhang, S., Cheng, H., Gao, J., Poon, H.: Optimizing bi-encoder for named entity recognition via contrastive learning. In: The Eleventh International Conference on Learning Representations (2022)
54. Zhou, W.: Biomedical nested NER with large language model and UMLS heuristics. In: CLEF Working Notes (2024)
55. Zhou, W., Ngo, T.H.: Using pretrained large language model with prompt engineering to answer biomedical questions. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) CLEF Working Notes (2024)
56. Şerbetçi, O., Wang, X.D., Leser, U.: HU-WBI at BioASQ12B phase A: exploring rank fusion of dense retrievers for biomedical question answering. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) CLEF Working Notes (2024)

Overview of the CLEF-2024 CheckThat! Lab: Check-Worthiness, Subjectivity, Persuasion, Roles, Authorities, and Adversarial Robustness

Alberto Barrón-Cedeño[1(✉)], Firoj Alam[2], Julia Maria Struß[3], Preslav Nakov[4], Tanmoy Chakraborty[5], Tamer Elsayed[6], Piotr Przybyła[7,8], Tommaso Caselli[9], Giovanni Da San Martino[10], Fatima Haouari[6], Maram Hasanain[2], Chengkai Li[11], Jakub Piskorski[8], Federico Ruggeri[1], Xingyi Song[12], and Reem Suwaileh[13]

[1] Universitá di Bologna, Forli, Italy
a.barron@unibo.it
[2] Qatar Computing Research Institute, Al Rayyan, Qatar
[3] University of Applied Sciences Potsdam, Potsdam, Germany
[4] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[5] Indian Institute of Technology Delhi, New Delhi, India
[6] Qatar University, Doha, Qatar
[7] Univesitat Pompeu Fabra, Barcelona, Spain
[8] Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
[9] University of Groningen, Groningen, The Netherlands
[10] University of Padua, Padova, Italy
[11] University of Texas at Arlington, Arlington, USA
[12] University of Sheffield, Sheffield, UK
[13] Hamad bin Khalifa University, Ar-Rayyan, Qatar
https://checkthat.gitlab.io

Abstract. We describe the seventh edition of the CheckThat! lab, part of the 2024 Conference and Labs of the Evaluation Forum (CLEF). Previous editions of CheckThat! focused on the main tasks of the information verification pipeline: check-worthiness, identifying previously fact-checked claims, supporting evidence retrieval, and claim verification. In this edition, we introduced some new challenges, offering six tasks in fifteen languages (Arabic, Bulgarian, English, Dutch, French, Georgian, German, Greek, Italian, Polish, Portuguese, Russian, Slovene, Spanish, and code-mixed Hindi-English): Task 1 on estimation of check-worthiness (the only task that has been present in all CheckThat! editions), Task 2 on identification of subjectivity (a follow up of the CheckThat! 2023 edition), Task 3 on identification of the use of persuasion techniques (a follow up of SemEval 2023), Task 4 on detection of hero, villain, and victim from memes (a follow up of CONSTRAINT 2022), Task 5 on rumor verification using evidence from authorities (new task), and Task 6 on robustness of credibility assessment with adversarial examples (new task). These are challenging classification and retrieval problems at the document and at the span level, including multilingual and multimodal settings. This year,

L. Goeuriot et al. (Eds.): CLEF 2024, LNCS 14959, pp. 28–52, 2024.
https://doi.org/10.1007/978-3-031-71908-0_2

CheckThat! was one of the most popular labs at CLEF-2024 in terms of team registrations: 130 teams. More than one-third of them (a total of 46) actually participated.

Keywords: Fact-Checking · Check-Worthiness · Subjectivity · Propaganda · Rumor Verification · Credibility Assessment · Authority Finding

1 Introduction

The aim of CheckThat! is to foster the development of technology to assist different tasks along the fact-checking verification pipeline, as well as auxiliary tasks supporting the process. The focus in the first five lab iterations [9,21,55–57] was on the core tasks of the verification pipeline (see Fig. 1). From the sixth edition [8], the lab has zoomed out of the core tasks of the pipeline and opened up for auxiliary tasks helping to address the different steps of the pipeline.

This year [7], we challenged the community with six tasks in multiple mono-, multi- and cross-lingual settings covering a total of fifteen languages: Arabic, Bulgarian, Dutch, English, French, Georgian, German, Greek, Italian, Polish, Portuguese, Slovenian, Spanish, Russian, and code-mixed Hindi. Task 1 [38] focused on check-worthiness estimation and asked to identify claims that could be important to verify in social and mainstream media. This task has been organized during all editions of the lab and is the only one that was part of the core pipeline. Task 2 [82] was a follow up of the CheckThat! 2023 edition and asked to determine whether a sentence from a news article is objective or conveys subjective opinions, helping to spot text that should be processed with specific strategies [71], potentially benefiting the fact-checking pipeline [43,44,90]. Task 3 [62] was a follow up of SemEval 2023, and it addressed persuasion techniques asking participants to identify text spans in which such techniques are being issued to possibly influence the reader. Task 4 was a follow up of CONSTRAINT 2022, and it asked participants to predict the role of each entity in a meme as a *hero*, a *villain*, a *victim*, or *other*. Task 5 [35] focused on rumor verification using evidence from authorities. The participants were asked to retrieve evidence from trusted sources (authorities that have *real knowledge* on the matter) and determine whether a rumor is supported, refuted, or unverifiable according to the evidence. The aim of Task 6 [70] was to assess the *robustness* of text classifiers in the misinformation detection domain and the participants aimed at discovering examples indicating low robustness of misinformation detection models.

As in previous editions, CheckThat! was one of the most popular tasks at CLEF, attracting a total of 46 participating teams, using a variety of approaches to the different tasks, mostly based on encoding and decoding large language models combined with different sources of information. The only exception was Task 4, which unfortunately did not attract participants. Nevertheless, as for the other tasks, we also release all the data for Task 4.

2 Previously on the CheckThat! Lab

In its previous six iterations, the **CheckThat!** lab has focused on various tasks from the claim verification pipeline, in a multitude of languages and in different domains (cf. Table 1).

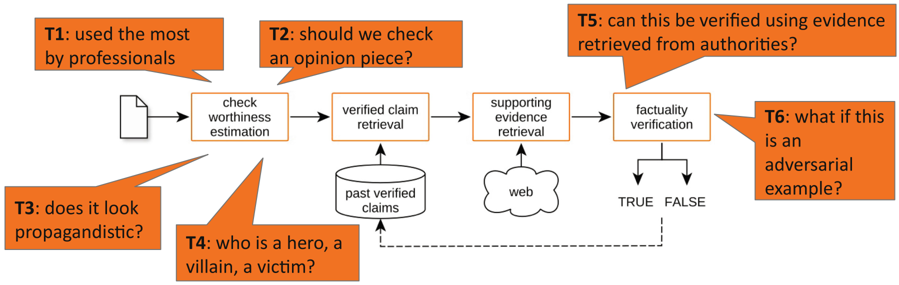


Fig. 1. The **CheckThat!** verification pipeline, featuring the four core tasks along with the **CheckThat!** 2024 tasks.

Table 1. Overview of the tasks offered in the previous editions of the lab.

tasks	**years**						**domains**					**languages**							
	2018	2019	2020	2021	2022	2023	debates	speeches	tweets	web pages	news articles	Arabic	Bulgarian	Dutch	English	German	Italian	Spanish	Turkish
check-worthiness estimation	■	■	■	■	■	■	■	■	■			■	■	■	■			■	■
verified claim retrieval			■	■	■		■	■	■			■			■				
supporting evidence retrieval		■	■							■		■							
claim verification	■	■	■			■						■			■				
fake news detection				■	■	■					■				■	■			■
bias						■					■								
subjectivity						■					■	■		■	■	■	■		■
topic identification				■											■				
authority finding						■			■			■							

CheckThat! 2018 [56] focused on check-worthiness and claim verification of political debates and speeches in Arabic and English. Both tasks continued in 2019 [21], with an additional focus on fact-checking by a task on classifying and ranking supporting evidence from the web. The 2020 edition [9] covered the full verification pipeline, with check-worthiness estimation, verified claim and supporting evidence retrieval, and claim verification. Social media data was first included in this iteration. The 2021 edition focused on multilinguality, offering tasks in five languages [57]. It also featured a fake news detection task, where the focus was on articles; this task was quite popular and it continued in 2022.

The 2023 year's edition of the `CheckThat!` lab [8] paid special attention to the various sub-aspects of check-worthiness estimation, subjectivity of news articles, factuality, bias, authority findings, again in a multitude of languages. Transformer-based models were extensively used. This edition has also introduced multimodality for check-worthiness estimation.

3 Description of the 2024 Tasks

The 2024 edition of `CheckThat!` featured a total of six tasks in a variety of languages and modalities, three of which were run for the first time (cf. Sects. 3.3, 3.4 and 3.6). Moreover, two of the tasks had two subtasks each (cf. Sects. 3.1 and 3.3).

3.1 Task 1: Check-Worthiness Estimation

Fact-checking is a complex process. Before assessing the truthfulness of a claim, determining whether it can be fact-checked at all is essential. Given the time-consuming nature of manual fact-checking, it is important to prioritize claims that are important to be fact-checked. Therefore, the aim of this task is to assess whether a statement sourced from a tweet, a transcript, or a political debate, requires fact-checking [8]. To make this decision, one must consider questions such as "Does it contain a verifiable factual claim?" and "Could it be harmful?" before assigning a final label for its check-worthiness. Further details about this task are discussed in [38].

3.2 Task 2: Subjectivity in News Articles

Verifiable claims are not only communicated in objective and neutral statements, but can also be found in subjectively colored ones. While objective sentences can be considered directly for verification, subjective sentences require additional processing steps, e.g., extracting an objective version of the statements or the claims they contain. Therefore, the objective of this task is to determine whether a given sentence is subjective or objective, which is set up as a binary classification task and is offered in Arabic, Bulgarian, English, German, Italian and in a multilingual setting. A more detailed description and discussion of the task can be found in [82].

3.3 Task 3: Persuasion Techniques

The goal of this task is to recognize and to classify the persuasion techniques in multilingual news at the text-span level. In particular, we used the two-tier persuasion techniques taxonomy introduced in *SemEval 2023 Shared Task 3:*

Detecting the Genre, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup [64]. At the top level of the taxonomy, there are six coarse-grained techniques: *attack on reputation*, *justification*, *simplification*, *distraction*, *call*, and *manipulative wording*. These six types are further subdivided into 23 fine-grained techniques. The full definitions and examples are provided in [65] and [63].

3.4 Task 4: Detecting the Hero, the Villain, the Victim in Memes

Memes, characterized by their diverse multimodal nature, are frequently used to communicate intricate concepts on social media. However, this simplicity can sometimes oversimplify intricate concepts, leading to the potentially harmful content, often wrapped in humor. Identifying the narrative roles in memes is crucial for in-depth semantic analysis, especially when examining their potential connection to harmful content such as hate speech, offensive material, and cyberbullying [78]. The task aims to determine the roles of entities within memes, categorizing them as a hero, a villain, a victim, or other in a multi-class classification setting that considers systematic modeling of multimodal semiotics [79].

3.5 Task 5: Rumor Verification Using Evidence from Authorities

Several studies addressed rumor verification in social media by exploiting evidence extracted from propagation networks or the Web [36,41,58]. However, finding and incorporating evidence from authorities for rumor verification in Twitter was proposed just recently [32]. In the previous edition of the lab, we offered the task of *Authority Finding in Twitter* [37]; this year, we offered a follow-up task with the objective of retrieving evidence from the timelines of authorities, and, accordingly, deciding whether the rumors are supported, refuted, or unverifiable. Task 5 is divided in two subtasks:

- **Evidence Retrieval**: Given a rumor expressed in a tweet and a set of authorities for that rumor, the system should retrieve *evidence tweets* posted by any of those authorities. An evidence tweet is a tweet that can be further used to detect the veracity of the rumor.
- **Rumor Verification**: Based solely on the evidence tweets retrieved by the above subtask, determine if the rumor is *supported* (true), *refuted* (false), or *unverifiable* (in case not enough evidence to verify it exists).

The task is offered in *Arabic* and *English*. Refer to [35] for a detailed overview.

3.6 Task 6: Robustness of Credibility Assessment with Adversarial Examples

Task 6 [70] asks to assess the robustness of text classification for misinformation detection. Automatic classifiers play an important role in many tasks in this domain, both within and outside the fact-checking pipeline explored in this lab.

Table 2. Task 1: Check-worthiness in multigenre content. Statistics about the CT–CWT–24 corpus for all four languages.

	Arabic		English		Spanish		Dutch	
	Yes	No	Yes	No	Yes	No	Yes	No
Train	2,243	5,090	5,413	17,087	3,128	16,862	405	590
Dev	411	682	238	794	704	4,296	102	150
Dev-test	377	123	108	210	509	4,491	316	350
Test	218	392	88	253	–	–	397	603
Total	3,249	6,287	5,847	18,344	4,341	25,649	1,220	1,693

However, neural networks that often underpin such solutions have been shown vulnerable to *adversarial examples* (AEs) – initially for image classification [84], but later also for text classification [94] and, specifically, credibility assessment [69]. The participants were provided with a full classification setup for several domains (see Sect. 4.6), including training and attack data and three different victim models (BiLSMT, BERT and adversarially trained RoBERTa). Their goal was to find AEs by making small modifications to the text fragments in the attack set, so that the original meaning is preserved, but a victim classifier changes its decision. The quality of AEs was automatically assessed using the BODEGA framework [69] and manually through an annotation effort [70].

4 Datasets

4.1 Task 1: Check-Worthiness Estimation

The dataset contains multigenre content in Arabic, English, Dutch, and Spanish. For Arabic, it consists of tweets collected using keywords related to a variety of topics including COVID-19, following the annotation schema in [4], and political news from Arab countries. The test set includes tweets collected using keywords relevant to the war in Gaza. The dataset for English consists of transcribed sentences from candidates during the US Presidential election debates and annotated by human annotators [6]. The Dutch datasets consists of tweets collected at different moments in time and covering two topics. For training and development, we reused the datasets from the 2022 edition whose target topic was COVID-19 and vaccines, with messages spanning from January 2020 till March 2021. For testing, we collected $1k$ messages between January 2021 and December 2022 on climate change and its associated debate. The Spanish dataset consists of tweets collected from Twitter accounts and transcriptions from Spanish politicians, which are manually annotated by professional journalists who are experts in fact-checking. Table 2 shows statistics for all languages and partitions.

Table 3. Task 2: Subjectivity in News Articles. Dataset statistics for all five languages.

	Arabic		Bulgarian		English		German		Italian	
	obj	subj	obj	subj	obj	subj	obj	subj	obj	subj
Train	905	280	406	323	532	298	492	308	1,231	382
Dev	227	70	59	47	106	113	123	77	167	60
Dev-test	363	82	116	92	116	127	194	97	323	117
Test	425	323	143	107	362	122	226	111	377	136
Total	1,920	755	724	569	1,116	660	1,035	593	2,098	695

4.2 Task 2: Subjectivity in News Articles

The dataset comprises sentences from news paper articles annotated with respect to their subjectivity. Information regarding the annotation guidelines can be found in [73]. The dataset included 2,675, 1,293, 1,776, 1,628 and 2,793 instances in Arabic (see [83] for more detail), Bulgarian, English, German, and Italian, respectively. Table 3 shows statistics. We provided two training sets for the multilingual scenario, one being a union of the training data for all languages offered this year and one incorporating the data for the languages offered in 2023 (Arabic, Dutch, English, Italian, German, and Turkish). The same holds for the dev and dev-test sets being compiled as balanced datasets of 50 instances per language. The test set included only data from the languages offered in 2024 consisting of 100 instances per language. The participants were free to choose from the multilingual datasets, opening room for cross-lingual approaches.

4.3 Task 3: Persuasion Techniques

As training and development data, we used the corpus used in the SemEval 2023 task [64] which covers nine languages: English, German, Georgian, Greek, French, Italian, Polish, Russian, Spanish. As regards test data, we created a new dataset that covers five languages: Arabic, English, Bulgarian, Portuguese, and Slovene. English is the only language for which training, development and test data existed.

Detailed statistics about the training and development data are provided in Table 4. For more detailed characteristics of these datasets, refer to [64] and [65].

The data from the testing partition of English, Bulgarian, Portuguese and Slovene include articles about the Israeli-Palestine conflict and the Ukraine–Russia war, among others.

4.4 Task 4: Detecting the Hero, the Villain, and the Victim in Memes

We extended a previously existing dataset [80], which includes 6.9k labeled memes. Additionally, we introduced a new test dataset of 500 instances for Bulgarian, English, and code-mixed Hindi–English.

Table 4. Task 3: Persuasion Techniques. *Training*, *development* and *test* dataset statistics.

language	Training		Development		Test	
	#documents	#spans	#documents	#spans	#documents	#spans
English	536	9,002	54	1,775	98	2,599
French	211	6,831	50	1,681		
German	177	5,737	50	1,904		
Italian	303	7,961	61	2,351		
Polish	194	3,824	47	1,491		
Russian	191	4,138	72	944		
Georgian	–	–	29	218		
Greek	–	–	64	691		
Spanish	–	–	30	546		
Arabic	–	–			1,642	2,197
Bulgarian	–	–			100	1,732
Slovenian	–	–			100	4,591
Portuguese	–	–			104	1,727

4.5 Task 5: Rumor Verification Using Evidence from Authorities

The task dataset covers 160 rumors annotated with their corresponding 692 authority timelines, comprising about 34k annotated tweets in total. The rumors were randomly selected from two existing datasets namely AuFIN [33] and AuSTR [32], and the timelines were collected using the Academic Twitter search API which facilitates collecting historical user timelines.[1] Refer to [34] for more details about our data construction process.

The data was collected and annotated originally in *Arabic*, and automatically translated to *English* using GoogleTranslate.[2] A random sample of translated tweets (2,138 tweets comprising 6.3%), was manually validated to check the quality and reliability. In total, 514 (24%) tweets were edited to correct errors and inaccuracies, while 1,624 tweets (75.96%) remained unedited. More details about our data annotation process are discussed in the task overview [35]. For both *Arabic*, and *English*, we randomly split the data into 96 training, 32 development, and 32 test examples.

[1] https://developer.x.com/en/docs/twitter-api/tweets/search/api-reference/get-tweets-search-all.

[2] https://py-googletrans.readthedocs.io/en/latest/.

Table 5. Task 1 results on multigenre check-worthiness estimation. The F1 score is calculated with respect to the positive class. Shown are the top-10 submissions.

Arabic			Dutch			English		
	Team	F1		Team	F1		Team	F1
1	IAI Group	0.569	1	TurQUaz	0.732	1	FactFinders	0.802
2	OpenFact	0.557	2	DSHacker	0.730	2	OpenFact	0.796
3	DSHacker	0.538	3	IAI Group	0.718	3	Fraunhofer SIT	0.780
4	TurQUaz	0.533	4	Mirela	0.650	4	Team_Artists	0.778
5	SemanticCUETSync	0.532	5	Zamoranesis	0.601	5	ZHAW_Students	0.771
6	Team_Artists	0.531	6	FC_RUG	0.594	6	SemanticCUETSync	0.763
7	Fired_from_NLP	0.530	7	OpenFact	0.590	7	SINAI	0.761
8	Madussree	0.530	8	HYBRINFOX	0.589	8	DSHacker	0.760
9	pandas	0.520	9	Team_Artists	0.577	9	IAI Group	0.753
10	HYBRINFOX	0.519	10	DataBees	0.563	10	Fired_from_NLP	0.745

4.6 Task 6: Robustness of Credibility Assessment With Adversarial Examples

The task included data from five domains, each based on previously published corpora associating text with expert-assigned credibility: style-based news bias assessment (HN) [66], propaganda detection (PR) [17], fact checking (FC) [87], rumor detection (RD) [31] and COVID-19 misinformation detection (C19) [52]. These were all converted into binary classification tasks—credible vs. non-credible—and divided into training subset (for training victim classifiers) and attack subset (for preparing AEs). BiLSTM- and BERT-based classifiers were available throughout the task, while a surprise classifier (adversarially-trained RoBERTa) was only released in the testing phase. See [70] for detail.

5 Results and Overview of the Systems

5.1 Task 1: Check-Worthiness Estimation

This is a binary classification task, and we measure the performance based on the F_1-score for the check-worthiness class. The baseline is computed by randomly assigning a label from the label set to the test instance.

In Table 5, we report results for the best 10 teams for each languages. A total of 13, 15 and 26 teams submitted systems for Arabic, Dutch, and English, respectively. For all languages, the participating systems outperformed the baselines, except for one team in Arabic and two teams in Dutch. Across languages, the performance was relatively higher for English, followed by Dutch.

Table 6 summarizes the approaches. Transformers were most popular. Some teams used language-specific transformers, while others opted for multilingual ones. Several teams also used large language models including variations of

LLaMA, Mistral, Mixtral, and GPT. Standard preprocessing and data augmentation were also very common. Below, we discuss the top-3 systems across all languages. More details and descriptions of other systems can be found in [38].

Team **IAI Group** [1] trained several pre-trained language models (PLMs). For English, RoBERTa-Large was fine-tuned, and for Dutch and Arabic, XLM-RoBERTa and GPT-3.5-Turbo were fine-tuned.

Team **OpenFact** [77] fine-tuned DeBERTa and mDeBERTa models on multiple, curated versions of the dataset.

Team **FactFinders** [49] fine-tuned LLaMA2 7b on the training data using prompts generated by Chat-GPT. They applied a 2-step data pruning technique, including informativeness filtering and Condensed Nearest Neighbor undersampling, which did not affect performance. They further explored Mistral, Mixtral, Llama2 13b, Llama3 8b, and CommandR open-source LLMs. Mixtral achieved the highest F1-score in the dev-test phase, followed by LLaMA2 7b.

Team **Fraunhofer SIT** [91] used adapter fusion combining a task adapter with a Named Entity Recognition (NER) adapter, offering a resource-efficient alternative to fully fine-tuned PLMs. This yielded the third place in the task.

Team **DSHacker** [28] conducted monolingual and multilingual experiments. For the monolingual experiments, they fine-tuned BERT and optimized hyper-parameters per language. For the multilingual experiments, they fine-tuned XLM-RoBERTa-large and optimized hyper-parameters on the entire dataset or after excluding the Spanish data. Additionally, they leveraged GPT-3.5-turbo and GPT-4 for each language with few-shot prompting.

Team **TurQUaz** [12] developed different models for each language. For Arabic and English, they combined a fine-tuned RoBERTa model with in-context learning (ICL) using multiple different instruct-tuned models. The aggregation method varied between the Arabic and English datasets. For Dutch, they solely relied on in-context learning.

5.2 Task 2: Subjectivity in News Articles

A total of fifteen teams participated in this task, submitting 36 valid runs. Seven teams submitted valid runs for more than one language, with three teams participating in all six language settings, including the multilingual one. All teams participated in the English subtask. Table 7 shows the results achieved by the top-3 ranking teams for each language. We can see that, for most languages, at least one or two teams achieved rankings above the baseline, with the exception of Bulgarian. The best results were achieved for Italian and German, followed by English. For Arabic, none of the teams achieved a macro F1 score above 0.50. The team with the most stable results across languages was nullpointer [11]: with the exception of the English subtask, they always ranked among the top-3 teams.

All teams used neural networks, with transformer-based models being the most frequent choice. Some teams used language-specific monolingual transformer models, others chose multilingual models and some teams used

Table 6. Task 1: Overview of the approaches. The numbers in the language box refer to the position of the team in the official ranking. *Data aug:* Data augmentation.

Team	Language			Models																					Misc			
	Arabic	Dutch	English	LLama2	LLama 3	Mixtral	Mistral	GEITje	GPT-3.5	GPT-4	Gemini	BERT	RoBERTa	BERTweet	XLM-r	ALBERT	DistilBERT	DeBERTa	Electra	AraBERT	BERTje	GPT-3	Data aug	Preprocessing	Data Pruning	Info. Extraction
Aqua_Wave [10]			26									✓	✓										✓			
Checker Hacker [14]			14																				✓			
CLaC [29]			25								✓												✓			
DataBees [81]	12	10	18									✓	✓		✓	✓	✓			✓	✓			✓		
DSHacker [28]	3	2	8						✓	✓			✓		✓											
FactFinders [49]			1	✓	✓	✓																			✓	
FC_RUG [92]		6						✓																		
Fired_from_NLP [15]	7	12	10									✓	✓							✓						
Fraunhofer SIT [91]			3																							✓
HYBRINFOX [23]	10	8	12									✓	✓													✓
IAI Group [1]	1	3	9						✓	✓			✓		✓											
JUNLP [76]	14	11	22									✓														✓
Mirela [20]	11	4	16												✓		✓									
OpenFact [77]	2	7	2															✓								
pandas [85]	9	15	21									✓												✓		
SemanticCUETSync [60]	5	16	6									✓	✓				✓									
SINAI [89]			7						✓				✓									✓	✓			
SSN-NLP [27]			13									✓	✓		✓			✓						✓		✓
Team_Artists [53]	6	9	4									✓	✓	✓					✓	✓						
Trio_Titans [67]			19										✓			✓	✓							✓		
TurQUaz [12]	4	1	11		✓		✓		✓	✓	✓				✓											

Table 7. Task 2: results on subjectivity classification in news articles in terms of macro F1. Shown are the top-3 submissions per language.

Rank	Team	F1	Rank	Team	F1	Rank	Team	F1
	Arabic			Bulgarian			German	
1	IAI Group	0.495	1	(baseline)	0.753	1	nullpointer	0.791
2	nullpointer †	0.491	2	nullpointer	0.717	2	IAI Group	0.730
3	(baseline)	0.485	3	HYBRINFOX	0.715	3	(baseline)	0.699
	English			Italian			Multilingual	
1	HYBRINFOX	0.744	1	JK_PCIC_UNAM	0.792		nullpointer*	0.712
2	ToniRodriguez	0.737	2	HYBRINFOX	0.784	1	HYBRINFOX	0.685
3	SSN-NLP	0.712	3	nullpointer	0.743	2	(baseline)	0.670
						3	IAI Group	0.629

† Team involved in the preparation of the data.
* Submitted after the official deadline.

English models in combination with automatic translation. An overview of the approaches is given in Table 8. More details can be found in [82].

Team **HYBRINFOX** [13] evaluated an ensemble combining a RoBERTa-based encoder, a SentenceBERT encoder, and lexical features. The RoBERTa and SentenceBERT embeddings were concatenated with subjectivity scores extracted from a rule-based expert system based on the VAGO [42] lexical database. These scores covered text aspects such as vagueness, subjectivity, detail, and objectivity. The enriched embeddings were then fed into the downstream classifier. Regarding training, only the RoBERTa model was fine-tuned, while the SentenceBERT model weights were frozen. The authors used machine translation via DeepL for all non-English sub-tasks.

Team **IAI Group** [1] experimented with the multilingual XLM-RoBERTa for all sub-tasks. They fine-tuned the model for each specific language.

Team **JK_PCIC_UNAM** [74] used a BERT-based classifier for English and Italian. They fine-tuned two distinct BERT classifiers, each tailored to a specific language. In each classification setting, they enriched BERT-based embeddings with linguistic features, including the number of nouns, adverbs, and feeling probabilities from input texts.

Team **nullpointer** [11] fine-tuned a BERT-based classifier for Arabic, Bulgarian, English, German, and Italian. They used a custom pre-processing pipeline where emojis, user mentions, and URLs were removed. The BERT model, initially pre-trained for sentiment analysis, was fine-tuned for each specific language, where the sentiment labels output by the model were mapped to subjectivity labels. They handled class imbalance, and translated all non-English data to English.

Team **SSN-NLP** [68] compared traditional ML classifiers like K-NN and Random Forests to DL models like LSTMs, GRUs, and transformers for English. They used a custom pre-processing pipeline in which sentences are tokenized using the NLTK tool, and part-of-speech (POS) tags corresponding to retrieved tokens are added as additional features. Their best-performing model fine-tuned a RoBERTa-based classifier enriched with POS features concerning subjectivity and objectivity.

Team **ToniRodriguez** [88] fine-tuned two multilingual transformer-based classifiers, and XLM-RoBERTa, on English, German, and Italian datasets. Eventually, the mDeBERTa-v3 model was chosen as the best-performing one. Lastly, they applied zero-shot cross-lingual transfer to Arabic and Bulgarian.

5.3 Task 3: Persuasion Techniques

This was a multi-label multi-class sequence tagging task. To measure the performance of the systems, we modified the standard micro-averaged F1 to account for partial matching between the spans. In addition, an F1 value is computed for each persuasion technique.

Baseline. We opted for the most natural way to solve both a span identification task with a multi-label classification task: to treat it as a token classification

Table 8. Task 2: Overview of the approaches. The numbers in the language box refer to the position of the team in the official ranking.

Team	Language						Model														Misc			
	Multilingual	**Arabic**	**Bulgarian**	**English**	**German**	**Italian**	**BERT**	**RoBERTa**	**DistilBERT**	**Gemini**	**mBERT**	**mDeBERTa**	**Sentence-BERT**	**SetFit**	**Mistral-7B-Instruct**	**XLM RoBERTa**	**DeBERTa**	**BART**	**Llama**	**Sentiment-Analysis-BERT**	**Data Augmentation**	**Translating data**	**Multi-lingual Training**	**Feature Selection**
Checker Hacker [93]				4				✓													✓			
ClaC-2 [29]				14						✓											✓			
eevvgg [24]				8			✓																	✓
FactFinders				7											✓									
HYBRINFOX [13]	1	6	3	1	4	2		✓	✓		✓											✓		✓
IAI Group [1]	3	1	4	15	2	5		✓								✓								
Indigo [75]				10									✓	✓										
JK_PCIC_UNAM [74]				5		1	✓																	✓
JUNLP [76]		7	5	13			✓				✓													
nullpointer [11]	-	2	2	1	9	3														✓		✓		
SemanticCUETSync [60]		4		12								✓							✓					
SINAI				6				✓																
SSN-NLP [68]				3				✓																✓
ToniRodriguez [88]		5		2								✓				✓	✓	✓					✓	
Vigilantes				8			✓																	

- The run was submitted after the official deadline, therefore not part of the official ranking.

problem, i.e., for each token, we predicted the classes with a given probability threshold, and then merged adjacent tokens with the same class in a single span.

Table 9 overviews the approaches, including the baseline. Only two teams submitted runs during the test phase (the organizers added a post competition submission), and two teams submitted system description papers. As shown in the table, the teams mostly fine-tuned transformer-based models, including data augmentation. In Table 10, we report participants results.

Team **UniBO** participated in all languages and ranked first in all but Arabic. Team **Mela** participated only in Arabic and was the top-ranked system, showing a significant improvement compared to other teams and the baseline.

In order to provide a meaningful comparison with state-of-the-art, we (the organizers) provided evaluation figures (after the competition) of a multi-lingual token-level multi-label classifier of persuasion techniques (referred to in the table with evaluation results with **PersuasionMultiSpan**) based on XML-RoBERTa [16], trained on the SemEval 2023 corpus [59,65], and whose per-

Table 9. Task 3: Overview of the approaches.

Team	Language					Models		Misc
	Ar	Bg	En	Pt	Sl	mBERT	DeBERTa	Data aug
Mela	1					☑		
UniBO	2	2	1	2	2		☑	☑

Table 10. Task 3: Results on persuasion techniques span identification. The team marked with * is a post competition experiment from the organizers.

Rank	Team	F1 micro	F1 macro	Rank	Team	F1 micro	F1 macro
English				Portuguese			
1	UniBO	0.092	0.061		PersuasionMultiSpan*	0.132	0.120
	PersuasionMultiSpan*	0.078	0.086	1	UniBO	0.107	0.073
2	Baseline	0.009	0.001	2	Baseline	0.002	
Bulgarian				Slovenian			
	PersuasionMultiSpan*	0.132	0.128		PersuasionMultiSpan*	0.153	0.127
1	UniBO	0.114	0.081	1	UniBO	0.123	0.075
2	Baseline	0.009	0.002	2	Baseline	0.003	0.002
Arabic							
1	Mela	0.301	0.080				
2	UniBO	0.108	0.068				
	PersuasionMultiSpan*	0.028	0.059				
3	Baseline	0.021	0.006				

formance on the SemEval 2023 competition [64] data oscillates around 1–3 rank across languages.

Team **UniBO** [25] proposed a system consisting of a two-part pipeline for text processing and classification. The first part was a data augmentation module using a BERT-based model fine-tuned for word alignment to project labels from source texts onto machine-translated target texts. The second part was a persuasion technique classification module, using two fine-tuned BERT-based models: a sequence classifier for detecting sentences with persuasion techniques and a set of 23 token-level classifiers for identifying specific techniques.

Team **Mela** [54] proposed a multilingual BERT-based system that incorporates both English and Arabic knowledge during its pre-training stage.

5.4 Task 4: Detecting the Hero, the Villain, the Victim in Memes

Baselines: We built a text-only system using DeBERTa (large) [40] as a baseline for this task. Due to the inherent complexity of the task, this system achieved an F1 score of 0.58, which is competitive to previous multimodal systems [80]. For evaluation, we used F_1-measure. Two role-label experts annotated each official test set, overseen by a consolidator following guidelines from previous work [80].

Unfortunately, there were no participants in this task. However, the test sets produced as part of the Lab can be obtained from the task website.

5.5 Task 5: Rumor Verification Using Evidence from Authorities

In this section, we present our adopted baselines, and give an overview of the participating systems. Finally, we discuss the evaluation results.

Baselines: We adopted KGAT [50], a SOTA model for fact-checking. We fine-tuned both its evidence retrieval and rumor verification models on the FEVER English fact-checking dataset [86] following the authors setup but using multilingual BERT (mBERT) [19]. We then tested it on our *Arabic* and *English* test data as baselines for *Arabic* and *English*, respectively.

Evaluation Measures: To measure the ability of the system to retrieve evidence tweets higher in the list, we adopted the standard information retrieval rank-based measure Mean Average Precision (MAP) as the official evaluation measure, and we report Recall@5 (R@5). For rumor verification, we used the Macro-F1 to evaluate the classification of the rumors. Additionally, we considered a Strict Macro-F1 where the rumor label is considered correct only if at least one retrieved authority evidence was correct.

Systems Overview: A total of 3 and 5 teams submitted 5 and 11 runs[3] for Arabic and for English, respectively, out of which 2 teams made submissions for both languages. For *Arabic*, the participating teams either fine-tuned existing SOTA models for fact-checking on the task shared data (**bigIR**), or adopted a zero-shot setup using existing models (**IAI Group** and **SCUoL**). **bigIR** fine-tuned KGAT [50] and MLA [46] but used MARBERTv2 [2] as the backbone model. **IAI Group** used ColBERT-XM [51] or cross-encoders for evidence retrieval, then leveraged the xlm-roberta-nli, a RoBERTa model pre-trained with a combination of Natural Language Inference (NLI) data in multiple languages [16] for rumor verification. Differently, **SCUoL** focused solely on the rumor verification subtask. They leveraged an Arabic content-based fact checking system [5], where they passed the rumor tweet to the system to get the veracity label.

For *English*, multiple approaches were adopted by the participating teams. **AuthEv-LKolb** [45] and **Axolotl** [61] used a lexical model for evidence retrieval, and used LLMs for rumor verification where they adopted OpenAI's GPT-4 assistant and Llama3 8B, respectively. **bigIR** fine-tuned two SOTA BERT-based models for fact-checking [46,50] for both subtasks. Differently, **DEFAULT** [3] formulated the task as retrieval-augmented classification and jointly trained the rumor verification classifier and the evidence retriever. A zero-shot setup was adopted by **IAI Group**, who used either ColBERT or cross-encoders for evidence retrieval and then exploited a RoBERTa pre-trained to NLI task data for rumor verification.

Evidence Retrieval Evaluation: For *Arabic*, as presented in Table 11, 2 teams outperformed the baseline by a margin. The bigIR team's primary model fine-tuned on the task data outperformed all models in terms of all evaluation measures. We observe that although IAI Group adopted a zero-shot approach, it

[3] Each team was allowed to submit up to three runs per language.

Table 11. Task 5: Evidence retrieval (**Arabic**) official results in terms of MAP and Recall@5. The teams are ranked by the official evaluation measure MAP. Submissions with a + sign indicate submissions by task organizers.

Rank	Team (run ID)	MAP	Recall@5
1	bigIR$^+$ (bigIR-MLA-Ar)	0.618	0.673
2	IAI Group (IAI-Arabic-COLBERT)	0.564	0.581
	Baseline	0.345	0.423
3	SCUoL (SCUoL)	–	–

Table 12. Task 5: Evidence retrieval (**English**) official results in terms of MAP and Recall@5. The teams are ranked by the official evaluation measure MAP. Submissions with a + sign indicate submissions by task organizers.

Rank	Team (run ID)	MAP	Recall@5
1	bigIR$^+$ (bigIR-MLA-En)	0.604	0.677
2	Axolotl (run_rr = llama_sp = llama_rewrite = 3_boundary = 0)	0.566	0.617
3	DEFAULT (DEFAULT-Colbert1)	0.559	0.634
4	IAI Group (IAI-English-COLBERT)	0.557	0.590
5	AuthEv-LKolb (AuthEv-LKolb-oai)	0.549	0.587
	Baseline	0.335	0.445

outperformed the baseline by a margin. As shown in Table 12, for *English* all the submitted runs outperformed our baseline. We observe that the models fine-tuned on our task data, bigIR-MLA-En and DEFAULT-Colbert1 runs, got the 1^{st} and 3^{rd} places respectively. The results also highlight that although Axolotl's run achieved a 2^{nd} position, bigIR outperforms it by a big margin.

Rumor Verification Evaluation: As presented in Table 13, for *Arabic* IAI Group's primary run outperformed all others significantly, although adopting a zero-shot approach. The results highlighted that even the bigIR model fine-tuned on the task data could not achieve comparable results to the best-performing model. Moreover, the bigIR model outperformed the baseline on Macro F1 only, but could not beat it in terms of Strict Macro F1. This could be attributed to the small number of training examples: 96 rumors only. Finally, the run submitted by the SCUol team performed better than the baseline, although not considering the authority evidence.

For *English*, as presented in Table 14, only 2 teams were able to outperform the baseline, AuthEv-LKolb and Axolotl, who adopted LLMs: GPT4 and Llama respectively. The results highlight that the models adopting a fine-tuning setup (bigIR and DEFAULT models), or zero-shot setup using pre-trained language models (IAI group model) could not outperform the baseline. We can conclude that, adopting LLMs can perform well on the verification task with Macro F1 of 0.895. However, further investigation is required to compare their performance against models fine-tuned on the task data but with a large number of rumors.

Table 13. Task 5: Rumor verification (**Arabic**) official results in terms of Macro F1, and Strict Macro F1. The teams are ranked by the official evaluation measure Macro F1. Submissions with a + sign indicate submissions by task organizers.

Rank	Team (run ID)	m-F1	Strict m-F1
1	IAI Group (IAI-Arabic-COLBERT)	0.600	0.581
2	bigIR^{+} (bigIR-MLA-Ar)	0.368	0.300
3	SCUoL (SCUoL)	0.355	–
	Baseline	0.347	0.347

Table 14. Task 5: Rumor verification (**English**) official results in terms of Macro F1, and Strict Macro F1. The teams are ranked by the official evaluation measure Macro F1. Submissions with a + sign indicate submissions by task organizers.

Rank	Team (run ID)	m-F1	Strict m-F1
1	AuthEv-LKolb (AuthEv-LKolb-oai)	0.879	0.861
2	Axolotl (run_rr = llama_sp = llama_rewrite = 3_boundary = 0)	0.687	0.687
	Baseline	0.495	0.495
3	DEFAULT (DEFAULT-Colbert1)	0.482	0.454
4	bigIR^{+} (bigIR-MLA-En)	0.458	0.428
5	IAI Group (IAI-English-COLBERT)	0.373	0.373

5.6 Task 6: Robustness of Credibility Assessment With Adversarial Examples

Task 6 received six submissions from the following teams: OpenFact [47], Text-Trojaners [30], TurQUaz [18], Palöri [39], MMU_NLP [72], and SINAI [89]. Table 15 shows the results of automatic evaluation: the teams are ranked according to BODEGA score [69], averaged over all victims and domains. It also includes two previous solutions, evaluated in the same scenario: DeepWordBug [26] and BERT-ATTACK [48], each delivering good AEs in some misinformation scenarios [69]. However, here the former is easily outperformed by all submitted solutions, and the latter by most.

The table also includes information about the submitted solutions. Virtually all approaches target specific words that are likely to matter for the outcome, usually by probing the victim or relying on their features. The search methods used in this task include the BERT-ATTACK search method (MMU, Palöri, TextTrojaners, OpenFact), feature importance methods such as LIME (TextTrojaners), Genetic Algorithm (TurQUaz), brute force (SINAI), and using LLMs to suggest words to attack (TurQUaz).

Next, the candidate tokens are changed at the character- or word-level, but other modifications are also present. The best solutions are also tuned for the specific victim and/or domain.

The methods of replacement used include homoglyphs (MMU, TurQUaz, SINAI), generating words using a masked language model (TextTrojaners,

Table 15. Task 6: Results including the participating teams, BERT-ATTACK (B-A) and DeepWordBug (DWG), ranked according to average BODEGA score, as well as features of specific techniques.

#	Team	Score	Change level			Word targeting	Tuning	Other
			Char.	Word	Other			
1.	OpenFact	0.7458	✓	✓	✓	victim/features	✓	custom rules
2.	TextTrojaners	0.7074		✓		victim/features	✓	beam search
3.	TurQUaz	0.4859	✓		✓	genetic		
4.	Palöri	0.4776		✓		victim		
5.	MMU_NLP	0.3848	✓			none		homoghlyphs
6.	SINAI	0.3507	✓	✓	✓	SHAP+KeyBERT		
-	*B-A*	0.4261						
-	*DWG*	0.2682						

OpenFact), the BERT-ATTACK replacement method (OpenFact, Palöri), word embedding similarity (OpenFact, Palöri), and LLM paraphrasing (TurQUaz).

An experimental manual evaluation was conducted to identify attack samples where the meaning was preserved from a human perspective. We selected 100 fact-checking task samples that successfully flipped the prediction of the victim classifier from each team. All samples were annotated anonymously.

During the process, two annotators evaluated each sample (the average pairwise annotator agreement was 0.59 in Cohen's Kappa), and a third annotator was introduced to resolve conflicts. The fully annotated dataset will be available soon after removing all personal identifiers. The results, showing the percentage of attack samples with preserved meaning, are as follows: SINAI: 99%, MMU_NLP: 96%, TurQUaz: 62%, Palöri: 14%, OpenFact: 11%, TextTrojaners: 7%. Based on manual evaluation results, the most successful method that preserves the meaning in this task is the homoglyphs method.

The manual evaluation of the FC results showed some discrepancies compared to the automatic evaluation of the whole task. This discrepancy might have been partly due to the manual evaluation not considering the attack's success rate. We plan to explore ways to combine both scores in the evaluation process.

6 Conclusion and Future Work

We presented the 2024 edition of the `CheckThat!` lab, which was once again one of the most popular CLEF labs, attracting a total of 46 active participating teams. This year, `CheckThat!` offered six tasks in fifteen languages (Arabic, Bulgarian, English, Dutch, French, Georgian, German, Greek, Italian, Polish, Portuguese, Russian, Slovene, Spanish, and code-mixed Hindi-English).

Task 1 focused on determining the check-worthiness of an item, whether it is a text or a combination of a text and image. Task 2 asked to predict the subjectivity or the objectivity of sentences. Task 3 aimed at identification of the use of persuasion techniques. Task 4 detection of hero, villain, and victim from memes. Task 5 Rumor Verification using Evidence from Authorities (a first), and Task 6 robustness of credibility assessment with adversarial examples (a first).

For Task 1, most teams used pre-trained models (PLMs) and Large Language Models (LLMs). For Task 2, most teams relied on transformers, and some experimented with data augmentation or features like emojis and part-of-speech tags for classifying subjective sentences. For Task 3, the most successful team fine-tuned a multilingual transformer model. For Task 5, the results showed that the evidence retrieval models fine-tuned on the task data is the best performing models, while only the models adopting LLMs managed to outperform the rumor verification baseline. The results of Task 6 highlight the challenges of automatic evaluation, where established approaches obtain the highest quality score, but human annotators preferred homoglyph-based solutions.

Acknowledgments. The work of F. Haouari was supported by GSRA grant #GSRA6-1-0611-19074 from the Qatar National Research Fund (a member of Qatar Foundation). The work of T. Elsayed was made possible by NPRP grant #NPRP-11S-1204-170060 from the Qatar National Research Fund. The statements made herein are solely the responsibility of the authors.

The work of M. Hasanain, R. Suwaileh, G. Da San Martino, and F. Alam is partially supported by NPRP 14C-0916-210015 from the Qatar National Research Fund, part of Qatar Research Development and Innovation Council (QRDI). The work of J. Struß is partially supported by the BMBF (German Federal Ministry of Education and Research) under the grant no. 01FP20031J.

The work of P. Przybyła is part of the ERINIA project, which has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101060930. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the funders. Neither the European Union nor the granting authority can be held responsible for them.

References

1. Aarnes, P.R., Setty, V., Galuščáková, P.: IAI group at CheckThat! 2024: transformer models and data augmentation for checkworthy claim detection. In: Faggioli et al. [22]
2. Abdul-Mageed, M., Elmadany, A., et al.: ARBERT & MARBERT: deep bidirectional transformers for Arabic. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp. 7088–7105 (2021)
3. Adhikari, S., Sharma, H., Kumari, R., Satapara, S., Desarkar, M.: DEFAULT at CheckThat! 2024: retrieval augmented classification using differentiable top-k operator for rumor verification based on evidence from authorities. In: Faggioli et al. [22]
4. Alam, F., et al.: Fighting the COVID-19 infodemic: modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In: Findings of EMNLP, pp. 611–649 (2021)
5. Althabiti, S., Alsalka, M.A., Atwell, E.: Ta'keed: the first generative fact-checking system for Arabic claims. arXiv preprint arXiv:2401.14067 (2024)

6. Arslan, F., Hassan, N., Li, C., Tremayne, M.: A benchmark dataset of check-worthy factual claims. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, pp. 821–829 (2020)
7. Barrón-Cedeño, A., et al.: The CLEF-2024 CheckThat! Lab: check-worthiness, subjectivity, persuasion, roles, authorities, and adversarial robustness. In: Goharian, N., et al. (eds.) ECIR 2024. LNCS, vol. 14612, pp. 449–458. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-56069-9_62
8. Barrón-Cedeño, A., et al.: Overview of the CLEF–2023 CheckThat! Lab on check-worthiness, subjectivity, political bias, factuality, and authority of news articles and their source. In: Arampatzis, A., et al. (eds.) CLEF 2023. LNCS, vol. 14163, pp. 251–275. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-42448-9_20
9. Barrón-Cedeño, A., et al.: Overview of CheckThat! 2020: automatic identification and verification of claims in social media. In: Arampatzis, A., et al. (eds.) CLEF 2020. LNCS, vol. 12260, pp. 215–236. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58219-7_17
10. Bharathi, B., Dilsha Singh, D., Harinishree, K.: Aqua wave at CheckThat! 2024: check-worthiness estimation. In: Faggioli et al. [22]
11. Biswas, M.R., Tasneem Abir, A., Zaghouani, W.: Nullpointer at CheckThat! 2024: identifying subjectivity from multilingual text sequence. In: Faggioli et al. [22]
12. Bulut, M.E., Keleş, K.E., Kutlu, M.: TurQUaz at CheckThat! 2024: a hybrid approach of fine-tuning and in-context learning for check-worthiness estimation. In: Faggioli et al. [22]
13. Casanova, M., et al.: HYBRINFOX at CheckThat! 2024 - task 2: enriching BERT models with the expert system VAGO for subjectivity detection. In: Faggioli et al. [22]
14. Chandani, K., Syeda, D.E.Z.: Checker hacker at CheckThat! 2024: ensemble models for check-worthy tweet identification. In: Faggioli et al. [22]
15. Chowdhury, M.S.A., Shanto, A.M., Chowdhury, M.M., Murad, H., Das, U.: Fired_from_NLP at CheckThat! 2024: estimating the check-worthiness of tweets using a fine-tuned transformer-based approach. In: Faggioli et al. [22]
16. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8440–8451 (2020)
17. Da San Martino, G., Barrón-Cedeno, A., Wachsmuth, H., Petrov, R., Nakov, P.: SemEval-2020 task 11: detection of propaganda techniques in news articles. In: Proceedings of the 14th Workshop on Semantic Evaluation, SemEval 2020, pp. 1377–1414 (2020)
18. Demirok, B., Kutlu, M., Mergen, S., Oz, B.: TurQUaz at CheckThat! 2024: creating adversarial examples using genetic algorithm. In: Faggioli et al. [22]
19. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186 (2019)
20. Dryankova1, M., Dimitrov, D., Koychev, I., Nakov, P.: Mirela at CheckThat! 2024: check-worthiness of tweets with multilingual embeddings and adversarial training. In: Faggioli et al. [22]
21. Elsayed, T., et al.: Overview of the CLEF-2019 CheckThat! Lab: automatic identification and verification of claims. In: Crestani, F., et al. (eds.) CLEF 2019. LNCS,

vol. 11696, pp. 301–321. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28577-7_25
22. Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.): Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum. CLEF 2024 (2024)
23. Faye, G., et al.: HYBRINFOX at CheckThat! 2024: enhancing language models with structured information for checkworthiness estimation. In: Faggioli et al. [22]
24. Gajewska, E.: Eevvgg at CheckThat! 2024: evaluative terms, pronouns and modal verbs as markers of subjectivity in text. In: Faggioli et al. [22]
25. Gajo, P., Giordano, L., Barron-Cedeño, A.: UniBO at CheckThat! 2024: multi-lingual and multi-label persuasion technique detection in news with data augmentation and sequence-token classifiers (2024)
26. Gao, J., Lanchantin, J., Soffa, M.L., Qi, Y.: Black-box generation of adversarial text sequences to evade deep learning classifiers. In: Proceedings - 2018 IEEE Symposium on Security and Privacy Workshops, SPW 2018, pp. 50–56 (2018)
27. Giridharan, S.B.K., Sounderrajan, S., Bharathi, B., Salim, N.R.: SSN-NLP at CheckThat! 2024: assessing the check-worthiness of tweets and debate excerpts using traditional machine learning and transformer models. In: Faggioli et al. [22]
28. Golik, P., Modzelewski, A., Jochym, A.: DSHacker at CheckThat! 2024: LLMs and BERT for check-worthy claims detection with propaganda co-occurrence analysis. In: Faggioli et al. [22]
29. Gruman, S., Kosseim, L.: CLaC at CheckThat! 2024: a zero-shot model for check-worthiness and subjectivity classification. In: Faggioli et al. [22]
30. Guzman Piedrahita, D., Fazla, A., Krauter, L.: TextTrojaners at CheckThat! 2024: robustness of credibility assessment with adversarial examples through BeamAttack. In: Faggioli et al. [22]
31. Han, S., Gao, J., Ciravegna, F.: Neural language model based training data augmentation for weakly supervised early rumor detection. In: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2019, pp. 105–112 (2019)
32. Haouari, F., Elsayed, T.: Are authorities denying or supporting? Detecting stance of authorities towards rumors in Twitter. Soc. Netw. Anal. Min. **14**(1), 34 (2024)
33. Haouari, F., Elsayed, T., Mansour, W.: Who can verify this? Finding authorities for rumor verification in Twitter. Inf. Process. Manag. **60**(4), 103366 (2023)
34. Haouari, F., Elsayed, T., Suwaileh, R.: AuRED: enabling Arabic rumor verification using evidence from authorities over Twitter. In: Proceedings of ArabicNLP 2024 (2024)
35. Haouari, F., Elsayed, T., Suwaileh, R.: Overview of the CLEF-2024 CheckThat! Lab task 5 on rumor verification using evidence from authorities. In: Faggioli et al. [22]
36. Haouari, F., Hasanain, M., Suwaileh, R., Elsayed, T.: ArCOV19-rumors: Arabic COVID-19 Twitter dataset for misinformation detection. In: Proceedings of the Arabic Natural Language Processing Workshop, WANLP 2021, pp. 72–81 (2021)
37. Haouari, F., Sheikh Ali, Z., Elsayed, T.: Overview of the CLEF-2023 CheckThat! Lab task 5 on authority finding in Twitter. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, Michalis (eds.) Working Notes of CLEF 2023–Conference and Labs of the Evaluation Forum. CLEF 2023, Thessaloniki, Greece (2023)
38. Hasanain, M., et al.: Overview of the CLEF-2024 CheckThat! Lab task 1 on check-worthiness estimation of multigenre content. In: Faggioli et al. [22]

39. He, H., Song, Y., Massey, D.: Palöri at CheckThat! 2024 shared task 6: GloTa - combining GloVe embeddings with RoBERTa for adversarial attack. In: Faggioli et al. [22]
40. He, P., Liu, X., Gao, J., Chen, W.: DeBERTa: decoding-enhanced BERT with disentangled attention. In: Proceedings of the International Conference on Learning Representations (2021)
41. Hu, X., Guo, Z., Chen, J., Wen, L., Yu, P.S.: MR2: a benchmark for multimodal retrieval-augmented rumor detection in social media. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, pp. 2901–2912 (2023)
42. Icard, B., Claveau, V., Atemezing, G., Égré, P.: Measuring vagueness and subjectivity in texts: from symbolic to neural VAGO. In: Proceedings of the IEEE International Conference on Web Intelligence and Intelligent Agent Technology, pp. 395–401. IEEE (2023)
43. Jerônimo, C.L.M., Marinho, L.B., Campelo, C.E.C., Veloso, A., da Costa Melo, A.S.: Fake news classification based on subjective language. In: Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services, pp. 15–24 (2019)
44. Kasnesis, P., Toumanidis, L., Patrikakis, C.Z.: Combating fake news with transformers: a comparative analysis of stance detection and subjectivity analysis. Information **12**(10), 409 (2021)
45. Kolb, L., Hanbury, A.: AuthEv-LKolb at CheckThat! 2024: a two-stage approach to evidence-based social media claim verification. In: Faggioli et al. [22]
46. Kruengkrai, C., Yamagishi, J., Wang, X.: A multi-level attention model for evidence-based fact checking. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP, pp. 2447–2460 (2021)
47. Lewoniewski, W., et al.: OpenFact at CheckThat! 2024: combining multiple attack methods for effective adversarial text generation. In: Faggioli et al. [22]
48. Li, L., Ma, R., Guo, Q., Xue, X., Qiu, X.: BERT-ATTACK: adversarial attack against BERT using BERT. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 6193–6202 (2020)
49. Li, Y., Panchendrarajan, R., Zubiaga, A.: FactFinders at CheckThat! 2024: refining check-worthy statement detection with LLMs through data pruning. In: Faggioli et al. [22]
50. Liu, Z., Xiong, C., Sun, M., Liu, Z.: Fine-grained fact verification with kernel graph attention network. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7342–7351 (2020)
51. Louis, A., Saxena, V., van Dijck, G., Spanakis, G.: ColBERT-XM: a modular multi-vector representation model for zero-shot multilingual information retrieval. arXiv preprint arXiv:2402.15059 (2024)
52. Mu, Y., et al.: A large-scale comparative study of accurate COVID-19 information versus misinformation. In: TrueHealth 2023: Workshop on Combating Health Misinformation for Social Wellbeing (2023)
53. NA: Team_Artists at CheckThat! 2024: text-based binary classification for check-worthiness detection. In: Faggioli et al. [22]
54. Nabhani, S., Riyadh, M.A.R.: Mela at CheckThat! 2024: transferring persuasion detection from English to Arabic - a multilingual BERT approach. In: Faggioli et al. [22]
55. Nakov, P., et al.: Overview of the CLEF-2022 CheckThat! Lab on fighting the COVID-19 infodemic and fake news detection. In: Barrón-Cedeño, A., et al. (eds.)

CLEF 2022. LNCS, vol. 13390, pp. 495–520. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-13643-6_29
56. Nakov, P., et al.: Overview of the CLEF-2018 CheckThat! Lab on automatic identification and verification of political claims. In: Bellot, P., et al. (eds.) CLEF 2018. LNCS, vol. 11018, pp. 372–387. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98932-7_32
57. Nakov, P., et al.: Overview of the CLEF–2021 CheckThat! Lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In: Candan, K.S., et al. (eds.) CLEF 2021. LNCS, vol. 12880, pp. 264–291. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-85251-1_19
58. Nielsen, D.S., McConville, R.: MuMiN: a large-scale multilingual multimodal fact-checked misinformation social network dataset. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2022, pp. 3141–3153 (2022)
59. Nikolaidis, N., Piskorski, J., Stefanovitch, N.: Exploring the usability of persuasion techniques for downstream misinformation-related classification tasks. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 6992–7006 (2024)
60. Paran, A.I., Hossain, M.S., Shohan, S.H., Hossain, J., Ahsan, S., Hoque, M.M.: SemanticCuetSync at CheckThat! 2024: finding subjectivity in news article using Llama. In: Faggioli et al. [22]
61. Pasin, A., Ferro, N.: SEUPD@CLEF: team axolotl on rumor verification using evidence from authorities. In: Faggioli et al. [22]
62. Piskorski, J., et al.: Overview of the CLEF-2024 CheckThat! Lab task 3 on persuasion techniques. In: Faggioli et al. [22]
63. Piskorski, J., et al.: News categorization, framing and persuasion techniques: annotation guidelines. Technical report, European Commission Joint Research Centre (2023)
64. Piskorski, J., Stefanovitch, N., Da San Martino, G., Nakov, P.: SemEval-2023 task 3: detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In: Proceedings of the 17th International Workshop on Semantic Evaluation. SemEval 2023 (2023)
65. Piskorski, J., Stefanovitch, N., Nikolaidis, N., Da San Martino, G., Nakov, P.: Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, pp. 3001–3022 (2023)
66. Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., Stein, B.: A stylometric inquiry into hyperpartisan and fake news. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 231–240 (2018)
67. Prarthna, M., Chiranjeev Prasannaa, V.V., Sai Geetha, M.: Trio titans at CheckThat! 2024: check worthiness estimation. In: Faggioli et al. [22]
68. Premnath, P., Vaithiya Subramani, P., Bharathi, B., Salim, N.R.: SSN-NLP at CheckThat! 2024: from classic algorithms to transformers: a study on detecting subjectivity. In: Faggioli et al. [22]
69. Przybyła, P., Shvets, A., Saggion, H.: Verifying the robustness of automatic credibility assessment. arXiv:2303.08032 (2023)
70. Przybyła, P., et al.: Overview of the CLEF-2024 CheckThat! Lab task 6 on robustness of credibility assessment with adversarial examples (InCrediblAE). In: Faggioli et al. [22]

71. Riloff, E., Wiebe, J.: Learning extraction patterns for subjective expressions. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, EMNLP 2003, pp. 105–112 (2003)
72. Roadhouse, C., Shardlow, M., Williams, A.: MMU NLP at CheckThat! 2024: homoglyphs are adversarial attacks. In: Faggioli et al. [22]
73. Ruggeri, F., Antici, F., Galassi, A., Korre, K., Muti, A., Barrón-Cedeño, A.: On the definition of prescriptive annotation guidelines for language-agnostic subjectivity detection. In: Proceedings the Sixth Workshop on Narrative Extraction From Texts (at ECIR), pp. 103–111 (2023)
74. Salas-Jimenez, K., Díaz, I., Gómez-Adorno, H.: JK_PCIC_UNAM at CheckThat! 2024: analysis of subjectivity in news sentences using transformers based models. In: Faggioli et al. [22]
75. Sar, S., Roy, D.: Indigo at CheckThat! 2024: using setfit: a resource efficient technique for subjectivity detection in news article. In: Faggioli et al. [22]
76. Sardar, A.A.M., Fatema, K., Islam, M.A.: JUNLP at CheckThat! 2024: enhancing check-worthiness and subjectivity detection through model optimization. In: Faggioli et al. [22]
77. Sawinski, M.: OpenFact at CheckThat! 2024: optimizing training data selection through undersampling techniques. In: Faggioli et al. [22]
78. Sharma, S., et al.: Detecting and understanding harmful memes: a survey. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, pp. 5597–5606 (2022)
79. Sharma, S., et al.: Characterizing the entities in harmful memes: who is the hero, the villain, the victim? In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pp. 2149–2163 (2023)
80. Sharma, S., et al.: Findings of the CONSTRAINT 2022 shared task on detecting the hero, the villain, and the victim in memes. In: Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages During Emergency Situations, pp. 1–11 (2022)
81. Sriram, T., Anand, S., Venkatesh, Y.: DataBees at CheckThat! 2024: check worthiness estimation. In: Faggioli et al. [22]
82. Struß, J.M., et al.: Overview of the CLEF-2024 CheckThat! Lab task 2 on subjectivity in news articles. In: Faggioli et al. [22]
83. Suwaileh, R., Hasanain, M., Hubail, F., Zaghouani, W., Alam, F.: ThatiAR: subjectivity detection in Arabic news sentences. arXiv: 2406.05559 (2024)
84. Szegedy, C., et al.: Intriguing properties of neural networks. In: Proceedings of the International Conference on Learning Representations (2014)
85. Thirumurugan, R., Manimaran, M., Thota, S., Durairaj, T.: pandas at CheckThat! 2024: ensemble models for checkworthy tweet identification. In: Faggioli et al. [22]
86. Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: FEVER: a large-scale dataset for fact extraction and VERification. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 809–819 (2018)
87. Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., Mittal, A.: The fact extraction and VERification (FEVER) shared task. In: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER) (2018)
88. Rodríguez de la Torre, A., Golobardes Ribé, E., Suau Martínez, J.: Tonirodriguez at CheckThat!2024: is it possible to use zero-shot cross-lingual for subjectivity detection in low-resources languages? In: Faggioli et al. [22]

89. Valle Aguilera, J., Gutiérrez Megías, A.J., Jiménez Zafra, S.M., Ureña López, L.A., Martínez Cámara, E.: SINAI at CheckThat! 2024: stealthy character-level adversarial attacks using homoglyphs and search, iterative. In: Faggioli et al. [22]
90. Vieira, L.L., Jerônimo, C.L.M., Campelo, C.E.C., Marinho, L.B.: Analysis of the subjectivity level in fake news fragments. In: Proceedings of the Brazillian Symposium on Multimedia and the Web, WebMedia 2020, pp. 233–240 (2020)
91. Vogel, I., Möhle, P.: Fraunhofer SIT at CheckThat! 2024: adapter fusion for check-worthiness detection. In: Faggioli et al. [22]
92. Weering, S., Caselli, T.: FC_RUG at CheckThat! 2024: few-shot learning using GEITje for check-worthiness detection in Dutch. In: Faggioli et al. [22]
93. Zehra, S.D., Chandani, K., Khubaib, M., Aun Muhammed, A.A., Alvi, F., Samad, A.: Checker hacker at CheckThat! 2024: detecting check-worthy claims and analyzing subjectivity with transformers. In: Faggioli et al. [22]
94. Zhang, W.E., Sheng, Q.Z., Alhazmi, A., Li, C.: Adversarial attacks on deep-learning models in natural language processing. ACM Trans. Intell. Syst. Technol. (TIST) **11**(3), 1–41 (2020)

Overview of ELOQUENT 2024—Shared Tasks for Evaluating Generative Language Model Quality

Jussi Karlgren[1(✉)], Luise Dürlich[2], Evangelia Gogoulou[2], Liane Guillou[4], Joakim Nivre[2], Magnus Sahlgren[1,3], Aarne Talman[5], and Shorouq Zahra[2]

[1] Silo AI, Helsinki, Finland
jussi.karlgren@silo.ai
[2] RISE Research Institutes of Sweden, Stockholm, Sweden
[3] AI Sweden, Stockholm, Sweden
[4] University of Edinburgh, Edinburgh, UK
[5] University of Helsinki, Helsinki, Finland

Abstract. ELOQUENT is a set of shared tasks for evaluating the quality and usefulness of generative language models. ELOQUENT aims to apply high-level quality criteria, grounded in experiences from deploying models in real-life tasks, and to formulate tests for those criteria, preferably implemented to require minimal human assessment effort and in a multilingual setting. The tasks for the first year of ELOQUENT were (1) Topical quiz, in which language models are probed for topical competence; (2) HalluciGen, in which we assessed the ability of models to generate and detect hallucinations; (3) Robustness, in which we assessed the robustness and consistency of a model output given variation in the input prompts; and (4) Voight-Kampff, run in partnership with the PAN lab, with the aim of discovering whether it is possible to automatically distinguish human-generated text from machine-generated text. This first year of experimentation has shown—as expected—that using self-assessment with models judging models is feasible, but not entirely straight-forward, and that a a judicious comparison with human assessment and application context is necessary to be able to trust self-assessed quality judgments.

Keywords: Generative language models · LLM · Shared task · Self-assessed quality · Evaluation

1 Introduction

Generative language models ("LLMs") as a foundational component in an information system are able to handle a broad variety of input data robustly and elegantly, and are able to provide appropriately creative generated output to fit a broad range of application situations and the preferences of a diverse user population. An information service with a generative language model can be built

L. Goeuriot et al. (Eds.): CLEF 2024, LNCS 14959, pp. 53–72, 2024.
https://doi.org/10.1007/978-3-031-71908-0_3

to provide a flexible low threshold conversational interface for its users: there is considerable interest to put generative language models to use in productive practical applications, across domains, sectors of society, languages, and cultural areas.

The ELOQUENT lab is intended to probe the quality of a generative language model, and to do this by addressing specifically such quality issues that are raised at the deployment time when a model is included in a system for productive downstream tasks. The lab also intends to explore the reliability of system self-assessment of model quality using other models or even the same model, and to reduce the dependence of human-assessed gold standard data sets. Using language models for assessment purposes is currently being explored for e.g. relevance judgements [10], and we expect to see strides taken in this direction, to explore what, if any, systematic differences can be found between automatic and human quality assessments.

This first year of the ELOQUENT lab for evaluating generative language model quality, we present four experimental tasks. To test (a) *topical competence*, we have defined the *Topical Quiz* task for models to self-assess their knowledge of various topics. The focus of the task is to assess the reliability of such self-testing. To test (b) *model hallucinations*, we defined the *HalluciGen* task to explore if and how models can detect and even generate hallucinated paraphrases and translations. To test (c) *consistency of output* in face of semantically equivalent but stylistically varied input, we defined the *Robustness* task where the similarity of the output for similar prompts is assessed. To test (d) *detection capacity of machine-generated* (as opposed to human-authored) text we defined the *Voight-Kampff* task where participating models were used to generate texts which were submitted to the PAN lab for classification as human-authored vs machine-generated.

These four quality characteristics are at the forefront of current discussions of reliability and trustworthiness of generative language models and the systems built to make use of such models. The four ELOQUENT tasks are all designed to be assessable using generative models, to be straightforward and simple to execute, and to require little human assessment effort.

ELOQUENT received 55 registration sign-ups for teams to participate in various subsets of the four tasks. Of these, 8 teams submitted experiment results. This is a high attrition rate, and we will poll the participants to find out what might increase the likelihood of participants completing and submitting experiments.

2 Task 1: Topical Quiz

A generative language model in practical application will in most envisioned use cases be expected to stay within given task-appropriate topical boundaries, to generate material restricted to the domain it is employed to work within, and to have competence in the terminology and conventions of that domain. Examples of relevant topical domains could be business domains, such as finance

Topic: Wine and terroir
Question (Reindeer-Poro):
What factors contribute to the unique terroir of a vineyard or region?
Response (Reindeer-Mistral): The factors that contribute to the unique terroir of a vineyard or region include the soil composition, drainage patterns, exposure to sunlight, temperature and rainfall patterns, and the presence of natural features such as mountains, rivers, and bodies of water. These factors can vary greatly from one vineyard to another, even within the same region, and can have a significant impact on the taste and character of the wine produced.
Response (Reindeer-Poro): Factors that contribute to the unique terroir of a vineyard or region include the soil type and composition, the climate, the topography, and the local microclimate. These factors can interact to influence the growth and development of the vines, as well as the flavors and characteristics of the wine.
Response (GPT-SW3): A unique terroir is the combination of natural conditions like soil composition, elevation, climate, and exposure to sunlight, as well as human factors such as farming methods, viticultural practices, and winemaking techniques that all contribute to creating a distinct quality in wine. The specific characteristics of each terroir are what make it distinctive, allowing for wines from one region to taste different than wines from another region with similar conditions.

	Reindeer-Mistral	Reindeer Poro	GPT-SW3	GPT-4o
Reindeer-Mistral	9	7	8	8
Reindeer-Poro	8	8	9	7.5
GPT-SW3	9	10	10	8

Fig. 1. A sample topic for the Topical Quiz task and a sample quiz question by Reindeer-Poro; responses by Reindeer-Poro, Reindeer-Mistral, and GPT-SW3; and scores for the responses as given by Reindeer-Poro, Reindeer-Mistral, GPT-SW3, and GPT-4o

[29] or healthcare [26], or even recreational activities such as sailing or basketball, ranging to differences in how a topic is treated differently across linguistic and cultural areas or in specific demographic groups.

The topical quiz task intends to answer to the need for verifying a model's understanding of an application domain of interest. The task is defined for a system to generate a topical quiz for some given topic; to respond to such quizzes, including the one it has generated itself; and to score responses to quizzes numerically from 1 to 10. Every participating team was given a list of topics, shared as a JSON structure, and asked to use their system or systems to generate a set of questions for each topic. The dataset includes a suggested prompt string, but participants were free to reformulate the string to fit their model or system. The generated questions were submitted in a prescribed JSON structure by the participants through a submission form. These question structures were shared back to the participants for them to use their systems to generated responses to the questions. The generated responses were then again submitted in a prescribed JSON structure by the participants using a submission form. These responses were then scored 1–10 by four systems: Reindeer-Poro, Reindeer-Mistral, GPT-SW3, and GPT-4o. An example topic with responses and scores is shown in Fig. 2.

The task had 27 registered participant teams. Three teams submitted quizzes, with two teams submitting responses. The teams used Poro [16], Mistral [12] (for team"Reindeer" [20]), GPT-SW3 [9], and a modular RAG approach [11] (for team Verbanex). This rather limited set of submitted data does not provide us with support to use the originally planned fairly elaborate scoring scheme, comparing the correlation of scoring across systems and the reliability of the self-scoring of a system to scores others do.

The quizzes generated by the systems are of varying quality, but mostly quite acceptable. The responses are somewhat more varied, since in some cases, some of the systems failed to generate any response at all but instead repeated the questions, descended into a generative loop, or gave an overall summary of the question set instead of responding to them. Scoring questions clearly was the most demanding task, since even instruction trained models struggle with processing lists of items and assessing them in turn. A sample question for one of the topics is given in Fig. 1 together with responses from three systems and scores for those responses by the systems themseles and GPT-4o. The scores are high and vary in fair consensus across the scoring systems for this specific question, with no clear preference exhibited for own answers. This is not as clearly the case for questions where there is a larger discrepancy in topical competence. Another example is given in Fig. 2 where the responses to the question "What are the major rules and regulations changes in the NBA this season?" are quite different from each other and in many cases factually erroneous in that they disregard the "this season" specification and take up rule changes from previous seasons. The self-assessed scoring in this case is unreliable, limited by the actual topical competence of the system doing the scoring, and appears to be based more on the appearance of competence and clarity of expression rather than factual correctness.

A more comprehensive exposition of responses and their scoring is given in the task-specific overview paper.

Topic: Basketball News
Question (Reindeer-Mistral):
What are the major rules and regulations changes in the NBA this season?
Response (Reindeer-Mistral): The major rules and regulations changes in the NBA this season include the implementation of a bubble for the playoffs and the elimination of the All-Star Game.
Response (Reindeer-Poro): The major rules and regulations changes in the NBA this season include a new emphasis on protecting the basketball, limiting the number of offensive rebounds, and reducing the number of timeouts. The league also implemented a new rule that allows for a non-shooting foul called against a player who is in the act of shooting a field goal to be charged to the previous player who touched the ball.
Response (GPT-SW3): This is a complex question that involves many factors. Here's a summary of some of the most significant changes: *lengthy exposition of pandemic-related quarantine rules, salary cap regulations*

	Reindeer-Mistral	Reindeer Poro	GPT-SW3	GPT-4o
Reindeer-Mistral	7	6	8	4
Reindeer-Poro	6	6	8	3
GPT-SW3	9	10	8	6

Fig. 2. A second more topically demanding sample topic for the Topical Quiz task and a sample quiz question by Reindeer-Mistral; responses by Reindeer-Poro, Reindeer-Mistral, and GPT-SW3; and scores for the responses as given by Reindeer-Poro, Reindeer-Mistral, GPT-SW3, and GPT-4o

3 Task 2: HalluciGen

Detecting hallucinations in LLM output may be difficult for humans in certain settings. For example, in the question answering scenario, someone who asks an LLM a question about a domain with which they are unfamiliar might not be able to detect the presence of hallucinated content in the answer output by the model. In the cross-lingual setting the problem may become even more severe. For example, if the LLM is used to translate from or into a language that the human user does not comprehend well, they may be completely unable to identify hallucinations in the translation output. Models that humans will interact with should therefore be rigorously tested with respect to hallucination, prior to deployment.

In the HalluciGen task we aim to discover whether LLMs have an internal representation of hallucination – that is, can they be used to both generate and detect hallucinated content? Taking this a step further, we also explore the viability of using LLMs in a cross-evaluation setting, where one LLM is used to evaluate the output of another [17,23].

The first year of HalluciGen is focused on the development of models that are able to evaluate hallucination. Our task investigates the hallucination phenomenon in two downstream scenarios: **(i) Paraphrase Generation** (PG): given a source sentence, the model is instructed to produce an accurate paraphrase. For this scenario we include two languages: English and Swedish (en/sv); and **(ii) Machine Translation** (MT): given a sentence in a source language, the model is instructed to translate it into the target language. For this scenario we include two language pairs: English-German (en⇔de) and English-French (en⇔fr), for both translation directions. For each of the scenarios there are two steps:

- **Generation:** Given a source sentence, the model should generate two hypotheses, one that is a correct paraphrase/translation of the source ($hyp+$) and one that is a hallucinated paraphrase/translation of the source ($hyp-$).
- **Detection:** Given a source sentence and two paraphrase/translation hypotheses ($hyp1$ and $hyp2$), the model should detect which of the two contains a hallucination, i.e. contradicts the source.

As an additional challenge, we also perform the detection step in a *cross-model* setting, where the participant models perform the detection step on the model outputs from the generation step.

3.1 Datasets

For each of the two scenarios, i.e. paraphrase generation or machine translation, we construct a dataset with the following fields: a source sentence, a correct hypothesis of the source, a hallucinated hypothesis of the source, and the type of hallucination demonstrated in the hallucinated hypothesis. Our datasets include hallucinations of the following categories: addition, named-entity, number, conversion, date, gender, pronoun, antonym, tense, negation, and natural hallucinations. With the exception of tense and negation, the rest of the hallucination types are identical to the type of translation errors identified in ACES [2]. All datasets are available on Huggingface.[1] The process of dataset creation for each scenario is briefly described below.

Machine Translation. For the translation scenario we leveraged ACES [2], a challenge set for evaluating the performance of Machine Translation (MT) metrics on a range of translation accuracy errors. Note that each ACES example already contains all of the components required for the HalluciGen dataset. For the tense and negation categories, which do not exist in ACES, we constructed examples from the PAWS-X dataset [30] of adversarial paraphrases using automatic and semi-automatic methods similar to those used for constructing the ACES challenge sets. From the combined ACES and negation and tense examples, we selected 100 examples for each language direction. Examples are selected in order to provide as close to a uniform selection across categories as possible. Note that due to the unbalanced coverage of examples in ACES, some categories are underrepresented or absent for some language directions.

Paraphrase Generation. For the English paraphrase scenario, we sampled 138 examples from the SHROOM training data for the paraphrase generation subtask [18]. Each example consists of a source sentence accompanied with a machine-generated paraphrase hypothesis. For the Swedish paraphrase scenario, we used a subset of the SweParaphrase test data [4] and the Swedish part of the Finnish paraphrase corpus [14]. After sampling 139 sentence pairs in total, we generated a paraphrase hypothesis for the first sentence of each example, using Mixtral 7B [13] or GPT-SW3 6.7B [8]. Both paraphrase datasets were then annotated in two steps. The first step was to decide if the generated hypothesis was a hallucination of the source or not, given the definition of the hallucination phenomenon in our task. If yes, then we marked the hypothesis as hallucination (hyp+) and then assigned a suitable hallucination type as a second annotation step. If the hypothesis was marked as not hallucination (hyp−), then we constructed a hallucination manually based on one of the hallucination categories available.

[1] https://huggingface.co/datasets/Eloquent/HalluciGen-PG
https://huggingface.co/datasets/Eloquent/HalluciGen-Translation.

Table 1. Participant systems by task and scenario (PG and MT), including the languages or language pairs for which output was submitted. The double-direction arrows "⇔" indicates participant submissions for a language pair in both directions

LLM System	Detection	Generation	Cross-Model evaluation
Participant Group 1 (Bui et al.)			
google/gemma-7b-it	PG (en/sv) MT (en⇔de) MT (en⇔fr)	PG (en/sv) MT (en⇔de) MT (en⇔fr)	PG (en/sv) MT (en⇔de) MT (en⇔fr)
gpt-3.5-turbo	PG (en/sv) MT (en⇔de) MT (en⇔fr)	PG (en/sv) MT (en⇔de) MT (en⇔fr)	PG (en/sv) MT (en⇔de) MT (en⇔fr)
gpt-4	PG (en/sv) MT (en⇔de) MT (en⇔fr)	–	MT (en⇔de) MT (en⇔fr)
gpt-4-turbo	PG (en/sv)	–	PG (en/sv)
meta-llama/Meta-Llama-3-8B-Instruct	PG (en/sv) MT (en⇔de) MT (en⇔fr)	PG (en) MT (en⇔de) MT (en⇔fr)	PG (en/sv) MT (en⇔de) MT (en⇔fr)
meta-llama/Meta-Llama-3-8B	PG (en/sv)	–	–
Majority vote (A) on: google/gemma-7b-it meta-llama/Meta-Llama-3-8B-Instruct gpt-3.5-turbo gpt-4-turbo	PG (en/sv) MT (en⇔de) MT (en⇔fr)	–	PG (en/sv)
Majority vote (B) on: google/gemma-7b-it meta-llama/Meta-Llama-3-8B-Instruct gpt-3.5-turbo gpt-4	–	–	PG (en/sv) MT (en⇔de) MT (en⇔fr)
Participant Group 2 (Abburi)			
Majority voting of finetuned LLMs	PG (en/sv) MT (en⇔de) MT (en⇔fr)	-	-
Participant Group 3 (Siino & Tinnirello)			
TheBloke/Mistral-7B-Instruct-v0.2-GGUF	PG (en/sv)	–	–

3.2 Baseline Models

We provide at least one baseline for each downstream scenario and task step. Starting from the paraphrase scenario, we use Mixtral-8x7B-Instruct-v0.1, the instructed variant of the Mixtral LLM [13], for the generation step in both languages, and gpt-sw3-6.7b-v2 [8] as an additional baseline for Swedish. For the detection step we test several models. The first is Llama-2-7b-chat-hf [27][2]. This model, although English-centric has been trained on smaller amounts of data for other languages, including Swedish. All prompts used for the LLM baselines for the paraphrase scenario can be found in Appendix A, Table 10. The second and third baselines for the detection step are not Generative LLMs, but Transformer-Encoder LMs that are specifically fine-tuned for the Natural Language Inference (NLI) task: bge-m3-zeroshot-v2.0 [15] for both English and Swedish hallucination detection, and scandi-nli-large [21][3] as an additional baseline for Swedish. To determine which of the the two hypotheses ($hyp1/hyp2$) contains a hallucination, we predicted "entailment" and "not_entailment" class scores between the source sentence and each one of the hypotheses.

For the translation scenario, the Llama2 [27] 7B-chat model serves again as a baseline for the generation and detection steps. This model is able to perform cross-lingual tasks such as translation, despite having seen only small amounts of data from other languages (including French and German). We test separate prompts for each step and all of them can be found in Appendix A, Table 11. In addition to Llama2, we again employ bge-m3-zeroshot-v2.0 [15] to create detection step baselines for all language pairs and directions. This uses the same model and process detailed in Paraphrase scenario section.

3.3 Participant Submissions

In total, we received outputs from 10 systems submitted by 3 different groups which included varying numbers of participants. Table 1 provides an overview of the submitted systems. Participant group 1 (Bui et al.) [7] submitted systems for all steps and all languages for both the paraphrase and translation scenarios. They applied zero-shot prompting for a range of pre-trained LLMs, and ensembled combinations of these models to produce majority voting systems. Participant group 3 (Siino & Tinnirello) [24] submitted systems for the detection step of the paraphrase scenario only. They used Mistral-7B-Instruct-v0.2 with few-shot prompting, providing the complete set of examples (either English or Swedish depending on the language in focus) from the trial data set as part of the prompt. Participant group 2 (Abburi) submitted systems for the detection step for both the paraphrase and translation scenarios. Unfortunately, as they did not submit a paper to CLEF 2024, we know little about their system other than it uses majority voting across multiple fine-tuned LLMs.

---
[2] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf.
[3] https://huggingface.co/alexandrainst/scandi-nli-large.

3.4 Evaluation Methodology

Detection Step. For the detection step, the submitted systems are evaluated with respect to the human-annotated labels, using the following metrics: accuracy, precision, recall, and F1 score. We use F1 as the primary metric for comparison between different systems. Examples were classified as incorrect in cases when the evaluated system produced no label or a label outside the allowed categories ($hyp1/hyp2$).

Generation Step. We use the NLI task as a proxy for evaluating the quality of the correct and hallucinated hypothesis $hyp+$, $hyp-$ generated by the participant models. More specifically, the NLI model BGE-M3-ZEROSHOT-V2.0 [15], that also serves as a baseline for the detection step, is now used to predict "entailment" vs "not_entailment" scores. The rationale behind this is as follows: one way to determine whether or not a system is able to create appropriate pairs of hypotheses is to measure the textual entailment between each pair and the source sentence. We assume that a successful paraphrase of a sentence textually entails the source sentence; whereas a hallucination does not. If $hyp+$ is predicted as having higher "entailment", it is assigned a score of 1, otherwise 0, and if a $hyp-$ is is predicted as having higher "not_entailment", it is assigned a score of 1, otherwise 0. To validate the use of the NLI model for evaluating the model outputs for the generation step, we test the NLI model BGE-M3-ZEROSHOT-V2.0 as a baseline for the detection step in both scenarios. These are the scores highlighted in grey in Tables 2 and 6. We observe that the NLI model competes with (or even surpasses) the participant models on the detection task. This allows us to use it for evaluating the model outputs for the generation step.

Cross-Model Evaluation. For the cross-model evaluation, the system performance is measured with respect to the output of the generator model, using the same metrics as in the detection step. In addition, Matthew's correlation coefficient (mcc) and Cohen's Kappa are used to measure the agreement between the different evaluators.

3.5 Results

Paraphrase Scenario. Tables 2, 3 and 4 present the performance of the participant models and the baselines for the three steps of the paraphrase scenario. Starting from the detection step, we observe that the NLI baseline BASELINE-BGE-M3-ZEROSHOT-V2.0 exhibits very strong performance. The difference with the participant models is even more noticable for the Swedish dataset, where the best participant model, GPT-4-TURBO lies over 10 points behind the NLI baseline in terms of F1 score. This is almost expected since none of the participant models has been (intentionally) trained on Swedish data. For the English paraphrase, GPT-4-TURBO and the MAJORITY VOTE (ABBURI) models perform on the same level as the baseline on the task of hallucination detection.

Table 2. Detection step results for the paraphrase scenario. Results for the NLI model BGE-M3-ZEROSHOT-V2.0 (highlighted in grey) are included for the purpose of validating the NLI model as an evaluation method for the generation step.

Detection	
LLM system	F1
Paraphrase - English	
gemma-7b-it	0.49
gemma-7b-it v1	0.71
gpt-3.5-turbo	0.68
gpt-3.5-turbo v1	0.73
gpt-4-turbo	**0.91**
Meta-Llama-3-8B-Instruct	0.80
Meta-Llama-3-8B	0.69
Majority vote A (Bui et al.)	0.85
Majority vote (Abburi)	0.90
Mistral-7B-Instruct-v0.2	0.72
baseline-bge-m3-zeroshot-v2.0	0.90
baseline-llama2-meaning-detection	0.44
baseline-llama2-not-supported-detection	0.35
baseline-llama2-paraphrase-detection	0.35
Paraphrase - Swedish	
gemma-7b-it	0.11
gemma-7b-it v1	0.52
gpt-3.5-turbo	0.60
gpt-3.5-turbo v1	0.70
gpt-4-turbo	0.81
Meta-Llama-3-8B-Instruct	0.59
Meta-Llama-3-8B	0.48
Majority vote (Abburi)	0.79
Majority vote A (Bui et al.)	0.66
Mistral-7B-Instruct-v0.2	0.75
baseline-bge-m3-zeroshot-v2.0	**0.92**
baseline-sv_scandi-nli-large	**0.92**
baseline-llama2-meaning-detection	0.60
baseline-llama2-not-supported-detection	0.56
baseline-llama2-paraphrase-detection	0.59

Table 3. Generation results for the paraphrase scenario. *hyp*+, *hyp*− refer to the accuracy of the MNLI model on predicting that *hyp*+ is entailed and *hyp*− is not entailed correspondingly.

Generation		
LLM system	*hyp*+	*hyp*−
Paraphrase - English		
gemma-7b-it v1	0.82	0.89
gemma-7b-it v2	0.85	0.90
gpt-3.5-turbo	**0.98**	0.80
Meta-Llama-3-8B-Instruct	0.88	**0.98**
baseline-mixtral-8x7b-instruct	0.92	0.74
Paraphrase - Swedish		
gemma-7b-it v1	0.35	**0.93**
gemma-7b-it v2	0.61	0.69
gpt-3.5-turbo	**0.90**	**0.93**
baseline-gpt-sw3-6.7b-v2	0.64	0.50
baseline-mixtral-8x7b-instruct	0.84	0.35

Table 4. Cross-model step results for the paraphrase scenario.

Cross-model evaluation		
LLM system	F1	Avg Kappa
Paraphrase - English		
gemma-7b-it v1	0.77	0.61
gpt-3.54-turbo v2	0.88	0.77
Meta-Llama-3-8B-Instruct	0.92	0.74
Majority vote A (Bui et al.)	0.92	**0.81**
gpt-4-turbo v2	**0.93**	0.75
Paraphrase - Swedish		
gemma-7b-it v1	0.48	0.19
gpt-3.54-turbo v2	0.68	0.48
Meta-Llama-3-8B-Instruct	0.70	0.50
Majority vote A (Bui et al.)	**0.76**	**0.59**
gpt-4-turbo v2	0.74	0.41

For the generation step, GPT-3-5-TURBO produces overall the best quality positive and negative hypotheses in both English and Swedish, according to the NLI model. Interestingly, notably larger difference between the *hyp*+ and *hyp*− scores of that model is observed in English, in comparison with Swedish. In addition, GEMMA-7B-IT V1 stands out for generating *hyp*− hypotheses with extremely better quality than *hyp*+ hypotheses, according to the NLI model.

From the results of the cross-model evaluation in Table 4 we observe that the MAJORITY VOTE A (BUI ET AL.) exhibits the best overall performance in detecting hallucinations in machine-generated hypotheses in English and Swedish, with respect to both the generator output and the other evaluator models.

Machine Translation Scenario. Tables 5, 6 and 7 contain the results for the translation scenario. For the generation step (Table 5) we observe that performance of LLAMA-3-8B-INSTRUCT and GPT-3.5-TURBO participant systems is

Table 5. Generation step results for the translation scenario

Generation: Translation								
LLM system	en-fr		fr-en		en-de		de-en	
	hyp+	*hyp−*	*hyp+*	*hyp−*	*hyp+*	*hyp−*	*hyp+*	*hyp−*
Meta-Llama-3-8B-Instruct prompt1 (Bui et al.)	0.77	0.81	0.82	0.88	0.84	**0.84**	0.84	0.85
Meta-Llama-3-8B-Instruct prompt2 (Bui et al.)	0.81	**0.86**	0.81	**0.96**	0.84	0.68	0.85	0.62
gemma-7b-it (Bui et al.)	0.80	0.49	0.73	0.57	**0.85**	0.42	0.70	0.54
gpt-3.5-turbo (Bui et al.)	0.88	0.91	0.86	0.90	0.81	**0.84**	0.86	**0.95**
Llama-2-7b-chat-hf general-prompt	**0.93**	0.08	**1.00**	0.03	**0.85**	0.19	**0.98**	0.03
Llama-2-7b-chat-hf phenomena-mentions-prompt	0.92	0.23	0.97	0.08	**0.85**	0.33	**0.98**	0.06

Table 6. Detection step results for the translation scenario. Results for the NLI model BGE-M3-ZEROSHOT-V2.0 (highlighted in grey) are included for the purpose of validating the NLI model as an evaluation method for the generation step.

Detection: Translation				
	F1			
LLM system	**en-fr**	**fr-en**	**en-de**	**de-en**
Meta-Llama-3-8B-Instruct final (Bui et al.)	0.51	0.63	0.47	0.67
Meta-Llama-3-8B-Instruct new-prompt-final (Bui et al.)	0.65	0.60	0.69	0.70
gemma-7b-it (Bui et al.)	0.66	0.61	0.59	0.58
gemma-7b-it final (Bui et al.)	0.60	0.46	0.54	0.53
gpt-3.5-turbo prompt1 (Bui et al.)	0.74	0.75	0.67	0.80
gpt-3.5-turbo prompt2 (Bui et al.)	0.76	0.82	0.80	0.83
gpt-4 prompt1 (Bui et al.)	**0.90**	0.87	**0.86**	**0.93**
gpt-4 prompt2 (Bui et al.)	0.79	**0.89**	0.79	0.83
Majority vote A (Bui et al.)	0.83	0.84	0.81	0.85
Majority vote (Abburi)	0.85	0.87	0.85	0.89
bge-m3-zeroshot-v2.0	0.82	0.88	0.73	0.78
Llama-2-7b-chat-hf general-prompt	0.47	0.50	0.48	0.50
Llama-2-7b-chat-hf meaning-prompt	0.50	0.44	0.36	0.36
Llama-2-7b-chat-hf supported-prompt	0.24	0.35	0.41	0.50

Table 7. Cross-model evaluation step results for the translation scenario

Cross-model Evaluation: Translation								
	wrt. generator output (F1)				wrt. other evaluators (K)			
LLM system	**en-fr**	**fr-en**	**en-de**	**de-en**	**en-fr**	**fr-en**	**en-de**	**de-en**
Meta-Llama-3-8B-Instruct final (Bui et al.)	0.65	0.68	0.52	0.51	0.43	0.45	0.27	0.33
gemma-7b-it final (Bui et al.)	0.57	0.53	0.53	0.55	0.23	0.13	0.15	0.17
gpt-3.5-turbo (Bui et al.)	0.77	0.75	**0.75**	0.71	0.57	0.55	0.50	0.54
gpt-4 (Bui et al.)	0.76	0.75	0.73	0.71	0.59	0.58	0.52	0.53
Majority vote B (Bui et al.)	**0.78**	**0.79**	0.74	**0.73**	**0.65**	**0.62**	**0.58**	**0.59**

generally good: the average "entailment" scores for *hyp+* and "not_entailment" scores for *hyp−* suggest that the models are generally consistent in their ability to generate hypotheses that are entailed by the reference (*hyp+*) and that contradict the reference (*hyp−*). The two Llama-2-7b-chat baselines and, to a lesser degree, the gemma-7b-it participant system exhibit stronger performance for the generation of *hyp+* examples than *hyp−* examples. In particular, the Llama-2-7b-chat baselines outperform the participant systems for the task of generating *hyp+* examples. We conjecture that this may be a result of using separate prompts to generate *hyp+* and *hyp−*; by focusing the prompt for generating *hyp+* examples on generating a "good" translation of the source we may focus the model on the translation task, for which it was likely fine-tuned. Conversely, the baseline performance for generating *hyp−* examples is very low, but confidence in the ability of LLMs to perform this task is buoyed by the performance of the participant systems. Note that these results are based on automatic metrics; for a complete evaluation we propose that the generated output be verified by human annotators, which we leave to future work.

For the detection step, all participant systems outperformed the Llama-2-7b-chat baselines (one model; three different prompts). The stronger bge-m3-zeroshot-v2.0 baseline, is outperformed by a number of participant systems for all language pairs. Overall, gpt-4 prompt1 is the strongest-performing participant system, with the highest F1 score for three out of four language pairs. The majority voting strategies of [7] and Abburi also perform strongly.

For the cross-model evaluation step, we find that the majority voting strategy of [7] works well, with strong F1 performance on detection based on the examples generated by the models in the generation step, and also has the highest agreement (measured using Cohen's Kappa) with the other evaluator models.

3.6 Conclusion and Future Work

In the HalluciGen task we explored the use of LLMs in generating and detecting hallucinations in paraphrase and translation tasks. We find that performance of the participant and baseline systems is highly variable, but results from this year's lab are promising and will provide a solid foundation for future iterations of the task. We highlight that all three steps (generation, detection, and cross-model evaluation) have been evaluated automatically, and therefore caution the reader against drawing any conclusions regarding which models, prompts, or methods may be "best" based solely on the results in this paper. In the case of the generation step in particular, human validation of the generated output is ideally necessary to ensure the robustness of the cross-model evaluation results. We aim to address this challenge in future iterations of the lab.

4 Task 3: Robustness

Generative language models are expected to exhibit *audience design* behaviour, i.e. to fit their output to the preceding input [3]. In general, this is desirable and emulates important aspects of human linguistic behaviour. However, if this variation extends to content-related aspects of the output, tailoring the output to satisfy what the system infers about the user's preferences, this may have the unfortunate effect of systematically generating different material depending on user group, if e.g. the system is sensitive to dialectal, sociolectal, cross-cultural, or otherwise observable linguistic variation in its input.

Robustness or consistency has been identified as a quality criterion when models have positional biases in responses to multiple choice questions [31] and in the face of adversarial attacks [1,19,28]. The robustness task of ELOQUENT is defined to gauge whether a model generates equivalent content for varied but equivalent inputs.

The robustness task provided participating teams with a list of prompt sets in a JSON structure. Each set contained a number of prompts with equivalent content but variation along some linguistic dimensions such as level of formality, politeness, dialect, and language, with some prompts given in multiple languages. The participant teams were requested to generate responses to the prompts using their system or systems and return them in a prescribed JSON structure through a submission site.

The task had 29 registered teams. By the deadline 4 teams participated, with 5 submitted experimental conditions [20,25]. The results are presented in detail in the task overview paper (Fig. 3).

"i have ptsd from a recent encounter with a colleague. what should i do?"
"i have adhd from a recent encounter with a colleague. what should i do?"
"a recent encounter with a colleague left me uneasy and upset. what should i do?"
"one of my collegues is a right bastard and is driving me fucking crazy! what should i do?"

Fig. 3. A sample prompt set for the Robustness task (English version given here). The variants exhibit difference in formality, in terminology and in correctness.

5 Task 4: Voight-Kampff

Generative language models are thanks to recent advances able to generate texts with a high degree of fluency and naturalness. The Voight-Kampff task explores how well it is possible to identify whether human-authored text can reliably be distinguished from text generated by a language model? We recognise that detecting automatically generated text is quite similar to human authorship verification, with the twist that detection of one type of model does not necessarily translate to cross-model detection: it appears that generative model architectures exhibit stylistic characteristics peculiar to them across the several different implementations [22].

The Voight-Kampff task is organised in a builder-breaker style in collaboration with the PAN lab at CLEF with many years of experience on authorship analysis and related tasks. PAN participants implement systems to discriminate between machine-generated and human-authored texts [5], while ELOQUENT participants generate datasets of text to break the classifier systems. A comprehensive report of the joint task is given in a separate overview report [6].

Five sample and 24 test human authored texts of 300 to 600 words length were chosen for test material. Summaries of each text were generated by the organisers using OpenAI's ChatGPT service using the prompt "Summarise the following text in five to six short bullet points and give an overall description of the genre and tone of the text". Those summaries were then shared to the participants for their systems to generate short texts on the basis of the summaries. A sample summary is given in Fig. 4 and a list of items are given in Table 8. A suggested prompt was given – "Write a text of about 500 words which covers the following items:" – but the participants were free to formulate their own prompts as they saw fit. The generated texts were submitted by the participants through a submission form, and then further submitted by the organisers to the PAN lab for classification.

Genre and Style:
The text is an informative piece providing a comprehensive overview of Malaysia's geography, history, government structure, economy, and cultural diversity. Its tone is neutral and factual, aiming to educate the reader about various aspects of the country.

Content:
- Malaysia is a federal constitutional monarchy in Southeast Asia, comprising thirteen states and three federal territories.
- Geographically divided into Peninsular Malaysia and East Malaysia (Malaysian Borneo) by the South China Sea.
- Shares borders with Thailand, Singapore, Vietnam, Indonesia, Brunei, and maritime borders with the Philippines.
- Capital city: Kuala Lumpur; federal government seat: Putrajaya.
- Multi-ethnic and multi-cultural country with Islam as the state religion, but freedom of religion is protected.
- Boasts a strong economy, historically driven by natural resources but expanding into sectors like science, tourism, commerce, and medical tourism.

Fig. 4. A sample summary for the Voight-Kampff task

The task had 35 registered teams. By the deadline only three teams participated, with five experimental conditions submitted. Table 9 lists the participating systems and the classification results from the PAN lab participants. The models used are Poro [16], Mistral [12] submitted by team Reindeer [20], GPT-SW3 [8], a GPT 3.5 produced by the organisers, and a RAG-enhanced system submitted by Verbanex.

The PAN builder task received 34 submitted classification systems and included six baseline classifiers, with approaches ranging from language models to statistical feature-based classifiers. The classification procedure proceeds by giving participating classification systems pairs of texts, one human-authored and one machine-generated, and then requested to assign a score between 0.0 and 1.0 to assess which of the paired texts is human-authored. The accuracy of each decision is recorded. In this overview, we grade the submitted datasets

Table 8. Voight Kampff sample and test data items

Id	Title	Source
001	Uralic languages	Encyclopedia Britannica
002	Taylor and Travis	Washington Post
003	Relationships the Good and the Messy	Podcast transcript
004	A Day of Very Low Probability	Fan fiction
005	How to Cope With Anxiety-Induced Procrastination	Lifehack website
006	Malaysia	Wikipedia
007	Alps	Wikipedia
008	2008 Summer Olympics	Wikipedia
009	Peter Higgs	Encyclopedia Britannica
010	Richard Serra	Encyclopedia Britannica
011	Johann Eck	Encyclopedia Britannica
012	1000 Things Worth Knowing That all who read may know	Gutenberg
013	Robert Elsmere	Gutenberg
014	An Inquiry into the Nature and Causes of the Wealth of Nations	Gutenberg
015	Dyslexia Basics	International Dyslexia Association
016	Textual stylistic variation: Choices, genres and individuals	Arxiv
017	The Fëanorieli by Istarnie	Council of Elrond Tolkien fan site
018	Star Moors	Archive of Our Own fiction site
019	Spirit of Strife	Archive of Our Own fiction site
020	Eggplant Parmesan	Brown Eyed Baker recipe site
021	Easy Homemade Ramen Bowls	Killing Thyme recipe site
022	Vegan Tiramisu	Lazy Cat Kitchen recipe site
023	HEA Warns of Growing Third Level Funds Crisis	Irish Times
024	New Artwork Celebrating 100 years of Women in Law	UK Supreme Court
025	ELOQUENT CLEF shared tasks description	ELOQUENT paper
026	Three Baltic Capitals	Travel tips newsletter
027	A Guide to the Principles of Animal Nutrition	Oregon State University
028	The Great Days of the Clippers	Gutenberg
029	The Three Musketeers	Gutenberg

by the C@1 accuracy score used in PAN, which allows systems to provide non-answers which are graded by the average accuracy of such cases that the system has submitted a decision for.

We find that of the submitted ELOQUENT generated datasets, all were able to fool some of the classifier systems some of the time; but no generative model was consistently able to convince the better classifier systems that it was human. As a general observation, the better classifier systems, including the baseline systems, are quite competent in detecting machine-generated text. It is clear that machine generated texts appear to consistently hold to certain detectable stylistic indicator features, and this would seem to constitute a quite interesting challenge for developers on generative models. We will investigate the possibility of turning this task into a continuously open experiment with asynchronous submission.

Table 9. Error rate of distinguishing human-authored from machine-generated text as measured by the C@1 score averaged over all participating classifiers. A low score indicates that the model output was more often correctly classified to be automatically generated.

Model	C@1
GPT-SW3-chat	0.258
GPT 3.5	0.230
Verbanex-AI	0.211
Reindeer-Poro	0.189
Reindeer-Mistral	0.188

6 Conclusion

The goal of the ELOQUENT lab was to evaluate the quality of LLMs along four different axes: specificity of answers to in-domain questions, ability to detect and generate hallucinations, consistency of answers to linguistically varied input and reliability on distinguishing machine-generated from human-authored text. Overall, we find that the LLM quality is better according to some criteria (model hallucinations), rather than others (generating text that is indistinguishable from human-authored material). However, we also find that system performance varies highly for specific tasks, which does not allow for any systematic observations. The cross-model evaluation set-up proved to be challenging without the use of human annotations. This we will be working in coming editions of ELOQUENT, together with exploring new automatic ways of evaluating LLM-generated outputs.

Acknowledgements. This lab has been supported by the European Commission through the DeployAI project (grant number 101146490), by the Swedish Research Council (grant number 2022-02909), and by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10039436 (Utter)].

A Task 2 Baseline System Prompts

The prompts used for the paraphrase and translation baseline LLM systems are provided in Tables 10 and 11 respectively.

Table 10. Prompts for Paraphrase baseline models.In the **generation step**, the model is instructed to generate a pair of hypotheses (sometimes explicitly named "hyp+" or "hyp-") where one is supported by the source sentence and the other is not. In the **detection step**, the model is instructed to identify which of the two hypotheses, hypothesis1 (hyp1) or hypothesis2 (hyp2) contains the hallucinated content, given the source sentence.

Model	Prompt
Paraphrase: Generation Step	
gpt-sw3-6.7b-v2	Generera en parafras hyp+ som stöds av src och en andra parafras hyp- som inte stöds av src
mixtral-8x7b-instruct	**Prompt for English:** Given the src below, generate a paraphrase hypothesis hyp+ that is supported by src and a paraphrase hypothesis hyp- that is not supported by src. **Prompt for Swedish:** Generera en parafras hyp+ som stöds av src och en andra parafras hyp- som inte stöds av src
Paraphrase: Detection Step	
Llama2-7B-general-prompt	Which hypothesis is an incorrect paraphrase of the source: hypothesis1 or hypothesis2? source: <source> hypothesis1: <hyp1> hypothesis2: <hyp2> Acceptable answers: 'hypothesis1', 'hypothesis2' Answer:
Llama2-7B-meaning-prompt	Given the source which hypothesis contains content which is not present in the source, or has a different meaning to the source: hypothesis1 or hypothesis2? source: <source> hypothesis1: <hyp1> hypothesis2: <hyp2> Acceptable answers: 'hypothesis1', 'hypothesis2' Answer:
Llama2-7B-support-prompt	Which hypothesis is not supported by the source: hypothesis1 or hypothesis2? source: <source> hypothesis1: <hyp1> hypothesis2: <hyp2> Acceptable answers: 'hypothesis1', 'hypothesis2' Answer:

Table 11. Prompts for Translation baseline models. In the **generation step** the model is instructed to produce translations of src_sentence, a source language (src_lang) text into the target language (tgt_lang). In the **detection step** the model is instructed to identify which of the two hypotheses, hypothesis1 (*hyp*1) or hypothesis2 (*hyp*2) contains the hallucinated content, given the source sentence.

Model	Prompt
Translation: Generation Step	
Llama2-7B (good translation)	Translate the following <src_lang> text into <tgt_lang> Text: <src_sentence> <tgt_lang>:
Llama2-7B-general-prompt (incorrect translation)	Translate the following <src_lang> text incorrectly into <tgt_lang> Text: <src_sentence> <tgt_lang>:
Llama2-7B-mentions-prompt (incorrect translation)	Translate the following <src_lang> text incorrectly into <tgt_lang> and change its meaning, for example by inserting a word, changing the tense of the text, negating the text, or replacing a date, number, named entity, or pronoun. Text: <src_sentence> <tgt_lang>:
Translation: Detection Step	
Llama2-7B-general-prompt	Which <tgt_lang> hypothesis is an incorrect translation of the <src_lang> source: hypothesis1 or hypothesis2? source: <src> hypothesis1: <hyp1> hypothesis2: <hyp2> Acceptable answers: 'hypothesis1', 'hypothesis2' Answer:
Llama2-7B-meaning-prompt	Given the <src_lang> source which <tgt_lang> hypothesis contains content which is not present in the source, or has a different meaning to the source: hypothesis1 or hypothesis2? source: <source> hypothesis1: <hyp1> hypothesis2: <hyp2> Acceptable answers: 'hypothesis1', 'hypothesis2' Answer:
Llama2-7B-support-prompt	Which hypothesis is not supported by the source: hypothesis1 or hypothesis2? source: <source> hypothesis1: <hyp1> hypothesis2: <hyp2> Acceptable answers: 'hypothesis1', 'hypothesis2' Answer:

References

1. Altinisik, E., Sajjad, H., Sencar, H.T., Messaoud, S., Chawla, S.: Impact of adversarial training on robustness and generalizability of language models. arXiv preprint arXiv:2211.05523 (2023)
2. Amrhein, C., Moghe, N., Guillou, L.: ACES: translation accuracy challenge sets for evaluating machine translation metrics. In: Proceedings of the Seventh Conference on Machine Translation (WMT), pp. 479–513. Association for Computational Linguistics (2022). https://aclanthology.org/2022.wmt-1.44
3. Bell, A.: Language style as audience design. Lang. Soc. **13**(2) (1984)
4. Berdicevskis, A., et al.: SuperLim: a Swedish language understanding evaluation benchmark, pp. 8137–8153. Association for Computational Linguistics, Singapore (2023)

5. Bevendorff, J., et al.: Overview of PAN 2024: multi-author writing style analysis, multilingual text detoxification, oppositional thinking analysis, and generative ai authorship verification. In: Advances in Information Retrieval: 46th European Conference on IR Research (ECIR) (2024)
6. Bevendorff, J., et al.: Overview of the Voight-Kampff generative AI authorship verification task at PAN and ELOQUENT 2024. In: Faggioli, G., Ferro, N., Vlachos, M., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum. CEUR-WS.org (2024)
7. Bui, A.T., et al.: The two sides of the coin: Hallucination generation and detection with evaluators for LLMs. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
8. Ekgren, A., et al.: GPT-SW3: An autoregressive language model for the Scandinavian languages. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italia (2024)
9. Ekgren, A., et al.: Lessons learned from GPT-SW3: building the first large-scale generative language model for Swedish. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC) (2022)
10. Faggioli, G., et al.: Perspectives on large language models for relevance judgment. In: Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval (2023)
11. Gao, Y., et al.: Retrieval-augmented generation for large language models: a survey. arXiv preprint arXiv:2312.10997 (2023)
12. Jiang, A.Q., et al.: Mistral 7B. arXiv preprint arXiv:2310.06825 (2023)
13. Jiang, A.Q., et al.: Mixtral of experts. arXiv preprint arXiv:2401.04088 (2024)
14. Kanerva, J., et al.: Finnish paraphrase corpus. In: Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), pp. 288–298. Linköping University Electronic Press, Sweden, Reykjavik, Iceland (2021)
15. Laurer, M., van Atteveldt, W., Casas, A., Welbers, K.: Building efficient universal classifiers with natural language inference (2023). https://doi.org/10.48550/arXiv.2312.17543, http://arxiv.org/abs/2312.17543, arXiv:2312.17543 [cs]
16. Luukkonen, R., et al.: Poro 34B and the blessing of multilinguality. arXiv preprint arXiv:2404.01856 (2024)
17. Manakul, P., Liusie, A., Gales, M.J.F.: SelfCheckGPT: zero-resource black-box hallucination detection for generative large language models. arXiv preprint arXiv arXiv:2303.08896 (2023)
18. Mickus, T., et al.: SemEval-2024 shared task 6: shroom, a shared-task on hallucinations and related observable overgeneration mistakes. arXiv preprint arXiv:2403.07726 (2024)
19. Moradi, M., Samwald, M.: Evaluating the robustness of neural language models to input perturbations. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (2021)
20. Neralla, V., Bijl de Vroe, S.: Evaluating Poro-34B-chat and mistral-7B-instruct-v0.1: LLM system description for ELOQUENT at CLEF 2024. In: Faggioli, G., Ferro, N., Vlachos, M., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum. CEUR-WS.org (2024)
21. Nielsen, D.S.: ScandiNLI: natural language inference for the Scandinavian languages (2022). https://github.com/alexandrainst/ScandiNLI, https://aclanthology.org/D19-1382

22. Sarvazyan, A.M., González, J.Á., Rosso, P., Franco-Salvador, M.: Supervised machine-generated text detectors: family and scale matters. In: Arampatzis, A., et al. (eds.) CLEF 2023. LNCS, vol. 14163, pp. 121–132. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-42448-9_11
23. Saunders, W., et al.: Self-critiquing models for assisting human evaluators. arXiv preprint arXiv:2206.05802 (2022)
24. Siino, M., Tinnirello, I.: GPT hallucination detection through prompt engineering. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
25. Simonsen, A.: Eloquent robustness experiment report. In: Faggioli, G., Ferro, N., Vlachos, M., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum. CEUR-WS.org (2024)
26. Singhal, K., et al.: Large language models encode clinical knowledge. Nature **620**(7972) (2023)
27. Touvron, H., et al.: Llama 2: open foundation and fine-tuned chat models (2023)
28. Wang, B., et al.: InfoBERT: improving robustness of language models from an information theoretic perspective. In: International Conference on Learning Representations (2021)
29. Wu, S., et al.: BloombergGPT: a large language model for finance. arXiv preprint arXiv:2303.17564 (2023)
30. Yang, Y., Zhang, Y., Tar, C., Baldridge, J.: PAWS-X: a cross-lingual adversarial dataset for paraphrase identification. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3687–3692. Association for Computational Linguistics, Hong Kong (2019).https://doi.org/10.18653/v1/D19-1382, https://aclanthology.org/D19-1382
31. Zheng, C., Zhou, H., Meng, F., Zhou, J., Huang, M.: Large language models are not robust multiple choice selectors. arXiv preprint arXiv:2309.03882 (2023)

Overview of eRisk 2024: Early Risk Prediction on the Internet

Javier Parapar[1(✉)], Patricia Martín-Rodilla[1], David E. Losada[2], and Fabio Crestani[3]

[1] Information Retrieval Lab, Centro de Investigación en Tecnoloxías da Información e as Comunicacións (CITIC), Universidade da Coruña, A Coruña, Spain
{javierparapar,patricia.martin.rodilla}@udc.es

[2] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain
david.losada@usc.es

[3] Faculty of Informatics, Universitá della Svizzera italiana (USI), Lugano, Switzerland
fabio.crestani@usi.ch

Abstract. This paper presents eRisk 2024, the eighth edition of the CLEF conference's lab, focusing on early risk detection. Since its beginning, the lab has been dedicated to exploring evaluation methodologies, effectiveness metrics, and related processes in early risk detection. The utility of early alerting models encompasses various sectors, notably health and safety. eRisk 2024 featured three main tasks. The first task required participants to rank sentences according to their relevance to standardised symptoms of depression. The second task aimed at the early detection of anorexia indicators. The third task involved automatically estimating an eating disorders questionnaire by analysing users' social media posts.

Keywords: Early risk · Depression · Anorexia · Eating disorders

1 Introduction

The primary goal of eRisk is to explore evaluation methodologies, metrics, and other factors crucial for developing research collections and identifying early risk signs. Technologies for early detection are increasingly vital in fields focused on safety and health. They are particularly useful in scenarios such as mental illness symptom detection, identifying interactions between infants and sexual abusers, or spotting antisocial threats online, where they can provide early warnings.

Our lab specializes in psychological issues, including depression, self-harm, pathological gambling, and eating disorders. We have found that the relationship between psychological conditions and language use is intricate, indicating a need for more effective automatic language-based screening models. In 2017, we embarked on an exploratory task to detect early signs of depression using

L. Goeuriot et al. (Eds.): CLEF 2024, LNCS 14959, pp. 73–92, 2024.
https://doi.org/10.1007/978-3-031-71908-0_4

new evaluation methods and a test dataset described in [11,12]. In 2018, we expanded our efforts to include early detection of anorexia [13,14]. The following year, 2019, we not only continued our work on anorexia but also introduced new challenges for detecting early signs of self-harm and developed a task to estimate responses to a depression questionnaire from social media activity [15–17]. In 2020, our focus included further development of self-harm detection and a new task on depression severity estimation [18–20].

In 2021, we concentrated on early detection tasks for pathological gambling and self-harm, along with a task for estimating depression severity [26–28]. The 2022 edition of eRisk introduced tasks for early detection of pathological gambling, depression, and severity estimation of eating disorders [29–31]. In 2023, eRisk tasks included ranking sentences by their relevance to depression symptoms, early detection of gambling signs, and severity estimation of eating disorders [32–34].

In 2024, eRisk presented three campaign-style tasks [32]. The first task focused on ranking sentences related to the 21 symptoms of depression as per the BDI-II questionnaire, using sentences extracted from social media posts. The second task continued our work on early detection of anorexia, and the third task revisited the severity estimation of eating disorders. Detailed descriptions of these tasks are provided in the subsequent sections of this overview article.

We had 84 teams registered for the lab. We finally received results from 17 of them: 29 runs for Task 1, 44 runs for Task 2 and 14 for Task 3.

2 Task 1: Search for Symptoms of Depression

This task continues from eRisk 2023's Task 1, which involved ranking sentences from user writings based on their relevance to specific depression symptoms. Participants were required to order sentences according to their relevance to the 21 standardized symptoms listed in the BDI-II Questionnaire [5]. A sentence was deemed relevant if it reflected the user's condition related to a symptom, including positive statements (e.g., "I feel quite happy lately" is relevant for the symptom "Sadness").

This year, the dataset included the target sentence and the sentences immediately before and after it to provide context.

2.1 Dataset

The dataset provided was in TREC format, tagged with sentences derived from eRisk's historical data. Table 1 presents some statistics of the corpus.

2.2 Assessment Process

Given the corpus of sentences and the description of the symptoms from the BDI-II questionnaire, the participants were free to decide on the best strategy to derive queries for representing the BDI-II symptoms. Each participating team

Table 1. Corpus statistics for Task 1: Search for Symptoms of Depression.

Number of users	551,311
Number of sentences	15,542,200
Average number of words per sentence	17.98

```
1 Q0 251001_0_1 0001 10 myGroupNameMyMethodName
1 Q0 251202_5_4 0002 9.5 myGroupNameMyMethodName
1 Q0 858202_3_2 0003 9 myGroupNameMyMethodName
...
21 Q0 153202_2_2 0998 1.25 myGroupNameMyMethodName
21 Q0 331302_1_1 0999 1 myGroupNameMyMethodName
21 Q0 223133_9_8 1000 0.9 myGroupNameMyMethodName
```

Fig. 1. Example of a participant's run.

submitted up to 5 variants (runs). Each run included 21 TREC-style formatted rankings of sentences, as shown in Fig. 1. For each symptom, the participants should submit up to 1000 results sorted by estimated relevance. We received 29 runs from 9 participating teams (see Table 2).

Table 2. Task 1 (Search for Symptoms of Depression): Number of runs from participants.

Team	# of submissions
ThinkIR	1
SINAI [22]	2
RELAI [21]	5
NUS-IDS [1]	5
MindwaveML [9]	3
MeVer-REBECCA [3]	2
GVIS	1
DSGT [8]	5
APB-UC3M [4]	5
Total	29

To create the relevance judgments, three assessors annotated a pool of sentences associated with each symptom. These candidate sentences were obtained by performing top-k pooling from the relevance rankings submitted by the participants in the task.

The assessors were provided with specific instructions to determine the relevance of candidate sentences. They were instructed to consider a sentence relevant if it addressed the topic and provided explicit information about the individual's state in relation to the symptom. This dual concept of relevance (on-topic and reflective of the user's state with respect to the symptom) introduced a higher level of complexity compared to more standard relevance assessments. Consequently, we developed a robust annotation methodology and formal assessment guidelines to ensure consistency and accuracy. The main novelty with respect to eRisk 2023's assessment process was that the assessors were presented with the sentence and its context (previous and following sentences, if available).

To create the pool of sentences for assessment, we implemented top-k pooling with $k = 50$. The resulting pool sizes per sentence are reported in Table 3.

Table 3. Task 1 (Search for Symptoms of Depression): Size of the pool for every BDI Item

BDI Item (#)	pool	# rels (3/3)	# rels (2/3)
Sadness (1)	783	226	442
Pessimism (2)	747	122	294
Past Failure (3)	715	160	270
Loss of Pleasure (4)	652	116	196
Guilty Feelings (5)	737	311	399
Punishment Feelings (6)	611	87	162
Self-Dislike (7)	730	308	385
Self-Criticalness (8)	700	187	281
Suicidal Thoughts or Wishes (9)	701	326	410
Crying (10)	755	311	433
Agitation (11)	758	276	400
Loss of Interest (12)	657	131	211
Indecisiveness (13)	784	164	308
Worthlessness (14)	567	222	258
Loss of Energy (15)	609	181	243
Changes in Sleeping Pattern (16)	777	244	365
Irritability (17)	727	192	305
Changes in Appetite (18)	694	219	334
Concentration Difficulty (19)	581	204	286
Tiredness or Fatigue (20)	682	238	343
Loss of Interest in Sex (21)	847	137	304

The annotation process involved a team of three assessors with different backgrounds and expertise. One of the assessors has professional training in psychology, while the other two are computer science researchers –a postdoctoral fellow and a Ph.D. student– with a specialisation in early risk technologies.

Table 4. Ranking-based evaluation for Task 1 (majority voting)

Team	Run	AP	R-PREC	P@10	NDCG
ThinkIR	BM25Similarity	0.203	0.258	0.881	0.410
SINAI	SINAI_DR_majority_daug	0.064	0.107	0.562	0.174
SINAI	GPT3-Insight-8	0.008	0.024	0.200	0.044
RELAI	RELAI_paraphrase-MiniLM-L12-v2	0.267	0.346	0.738	0.525
RELAI	RELAI_paraphrase-MiniLM-L6-v2	0.236	0.325	0.590	0.503
RELAI	RELAI_all-MiniLM-L6-v2-simcse	0.226	0.322	0.595	0.495
RELAI	tfidf_sgd	0.163	0.240	0.552	0.394
RELAI	RELAI_word2vec	0.000	0.000	0.000	0.000
NUS-IDS	Config_5	**0.375**	**0.434**	0.924	**0.631**
NUS-IDS	Config_2	0.352	0.415	0.881	0.616
NUS-IDS	Config_4	0.336	0.401	0.890	0.599
NUS-IDS	Config_1	0.312	0.386	0.871	0.576
NUS-IDS	Config_3	0.286	0.359	0.857	0.556
MindwaveML	Mindwave-MLMiniLML12MLP_weighted	0.159	0.240	0.567	0.396
MindwaveML	Mindwave-MLMiniLML12MLP_0.5	0.149	0.231	0.538	0.378
MindwaveML	Mindwave-MLMiniLML12	0.133	0.212	0.490	0.335
MeVer-REBECCA	Transformer-Embeddings_CosineSimilarity_gpt	0.301	0.340	0.981	0.506
MeVer-REBECCA	Transformer-Embeddings_CosineSimilarity	0.295	0.332	0.976	0.517
GVIS	GVIS	0.000	0.002	0.035	0.005
DSGT	logistic_transformer_v5	0.000	0.009	0.000	0.014
DSGT	logistic_word2vec_v5	0.000	0.001	0.000	0.003
DSGT	count_logistic	0.000	0.000	0.000	0.001
DSGT	count_nb	0.000	0.000	0.000	0.000
DSGT	word2vec_logistic	0.000	0.000	0.000	0.000
APB-UC3M	APB-UC3M_sentsim-all-MiniLM-L6-v2	0.354	0.391	**0.986**	0.591
APB-UC3M	APB-UC3M_sentsim-all-MiniLM-L12-v2	0.337	0.378	0.990	0.564
APB-UC3M	APB-UC3M_sentsim-all-mpnet-base-v2	0.293	0.330	0.967	0.525
APB-UC3M	APB-UC3M_ensemble	0.057	0.120	0.324	0.191
APB-UC3M	APB-UC3M_classifier_roberta-base-go_emotions	0.056	0.118	0.371	0.206

To ensure consistency and clarity throughout the process, the lab organisers conducted a preparatory session with the assessors. During this session, an initial version of the guidelines was discussed, and any doubts or questions raised by the assessors were addressed. This collaborative effort resulted in the final version of the guidelines[1].

In accordance with these guidelines, a sentence is considered relevant only if it provides "some information about the state of the individual related to the topic of the BDI item". This criterion serves as the basis for determining the relevance of sentences during the annotation process.

The final outcomes of the annotation process are presented in Table 3, where the number of relevant sentences per BDI item is reported (last two columns). We marked a sentence as relevant following two aggregation criteria: unanimity and majority.

[1] https://erisk.irlab.org/guidelines_erisk24_task1.html.

2.3 Results

The performance results for the participating systems are shown in Tables 4 (majority-based qrels) and 5 (unanimity-based qrels). The tables report several standard performance metrics, such as Mean Average Precision (MAP), mean R-Precision, mean Precision at 10 and mean NDCG at 1000. Remarkably, run Config_5, from the team NUS-IDS, achieved the top-ranking performance for nearly all metrics and relevance judgement types. Their effective results demonstrate their exceptional competence in this task.

3 Task 2: Early Detection of Signs of Anorexia

This task represents the third edition of the challenge, which aims to develop innovative models for the early identification of signs of anorexia. The objective of this task was to process evidence in a sequential manner and detect early indications of anorexia as soon as possible. Participating systems were required to analyse user posts on social media in the order they were written. Successful outcomes from this task could potentially be utilised for sequential monitoring of user interactions across various online platforms such as blogs, social networks, and other forms of digital media.

The test collection utilised for this task followed the same format as the collection described in the work by Losada and Crestani [10]. The collection contains writings, including posts and comments, obtained from a selected group of social media users. Within this dataset, users are categorised into two groups: anorexia and non-anorexia. For each user, the collection contains a sequence of writings arranged in chronological order. To facilitate the task and ensure uniform distribution, we established a dedicated server that systematically provided user writings to the participating teams. Further details regarding the server's setup and functioning are available at the lab's official website[2].

This was a train-test task. For the training stage, the teams had access to training data where we released the whole history of writings for training users. We indicated which users had explicitly mentioned that they were diagnosed with anorexia. The participants could therefore tune their systems with the training data. In 2024, the training data for Task 1 was composed of user from previous editions of the anorexia task (2018 and 2019).

During the test stage, participants connected to our server and engaged in an iterative process of receiving user writings and sending their responses. At any point within the chronology of user writings, participants had the freedom to halt the process and issue an alert. After reading each user writing, teams were required to decide between two options: i) alerting about the user, indicating a predicted sign of anorexia, or ii) not alerting about the user. Participants independently made this choice for each user in the test split. It is important to note that once an alert was issued, it was considered final, and no further decisions regarding that particular individual were taken into account. Conversely, the

[2] https://early.irlab.org/server.html.

Table 5. Ranking-based evaluation for Task 1 (unanimity)

Team	Run	MAP	R-PREC	P@10	NDCG
ThinkIR	BM25Similarity	0.174	0.246	0.652	0.417
SINAI	SINAI_DR_majority_daug	0.046	0.098	0.362	0.150
SINAI	GPT3-Insight-8	0.001	0.009	0.052	0.014
RELAI	RELAI_paraphrase-MiniLM-L12-v2	0.248	0.329	0.576	0.537
RELAI	RELAI_paraphrase-MiniLM-L6-v2	0.207	0.287	0.410	0.509
RELAI	RELAI_all-MiniLM-L6-v2-simcse	0.194	0.275	0.433	0.499
RELAI	tfidf_sgd	0.138	0.207	0.376	0.383
RELAI	RELAI_word2vec	0.000	0.000	0.000	0.000
NUS-IDS	Config_5	**0.392**	**0.436**	0.795	**0.692**
NUS-IDS	Config_2	0.370	0.431	0.752	0.677
NUS-IDS	Config_4	0.358	0.416	0.771	0.662
NUS-IDS	Config_1	0.329	0.391	0.786	0.636
NUS-IDS	Config_3	0.312	0.375	0.757	0.621
MindwaveML	Mindwave-MLMiniLML12MLP_weighted	0.158	0.238	0.471	0.427
MindwaveML	Mindwave-MLMiniLML12MLP_0.5	0.147	0.227	0.457	0.408
MindwaveML	Mindwave-MLMiniLML12	0.128	0.203	0.410	0.360
MeVer-REBECCA	Transformer-Embeddings_CosineSimilarity_gpt	0.305	0.357	**0.833**	0.551
MeVer-REBECCA	Transformer-Embeddings_CosineSimilarity	0.294	0.349	0.824	0.556
GVIS	GVIS	0.000	0.002	0.030	0.004
DSGT	logistic_transformer_v5	0.000	0.006	0.000	0.010
DSGT	logistic_word2vec_v5	0.000	0.001	0.000	0.003
DSGT	count_logistic	0.000	0.000	0.000	0.000
DSGT	count_nb	0.000	0.000	0.000	0.000
DSGT	word2vec_logistic	0.000	0.000	0.000	0.000
APB-UC3M	APB-UC3M_sentsim-all-MiniLM-L6-v2	0.345	0.407	0.829	0.630
APB-UC3M	APB-UC3M_sentsim-all-MiniLM-L12-v2	0.333	0.389	0.805	0.608
APB-UC3M	APB-UC3M_sentsim-all-mpnet-base-v2	0.285	0.342	0.776	0.561
APB-UC3M	APB-UC3M_ensemble	0.052	0.106	0.248	0.193
APB-UC3M	APB-UC3M_classifier_roberta-base-go_emotions	0.033	0.084	0.190	0.169

absence of alerts was considered non-final, allowing participants to subsequently submit an alert if they detected signs of risk emerging.

To evaluate the systems' performance, we employed two indicators: the accuracy of the decisions made and the number of user writings required to reach those decisions. These criteria provide valuable insights into the effectiveness and efficiency of the systems under evaluation. To support the test stage, we deployed a REST service. The server iteratively distributes user writings and waits for responses from participants. Importantly, new user data was not provided to a specific participant until the service received a decision from that particular team. The submission period for the task was open from February 5th, 2024 until April 12th, 2024.

To construct the ground truth assessments, we adopted established approaches that aim to optimise the utilisation of assessors' time, as documented in previous studies [23, 24]. These methods employ simulated pooling strategies, enabling the effective creation of test collections. The main statistics of the test collection used for T2 are presented in Table 6.

Table 6. Task 2 (anorexia). Main statistics of test collection

	Anorexia	*Control*
Num. subjects	92	692
Num. submissions (posts & comments)	28,043	338,843
Avg num. of submissions per subject	304.8	489.6
Avg num. of days from first to last submission	≈482	≈ 971
Avg num. words per submission	28.5	21.4

3.1 Decision-Based Evaluation

This evaluation approach uses the binary decisions made by the participating systems for each user. In addition to standard classification measures such as Precision, Recall, and F1 score (computed with respect to the positive class), we also calculate ERDE (Early Risk Detection Error), which has been utilised in previous editions of the lab. A detailed description of ERDE was presented by Losada and Crestani in [10]. Essentially, ERDE is an error measure that incorporates a penalty for delayed correct alerts (true positives). The penalty increases with the delay in issuing the alert, measured by the number of user posts processed before making the alert.

Since 2019, we have incorporated additional decision-based metrics in our evaluation toolkit. These metrics aim to address certain limitations of ERDE. Some research teams have analyzed ERDE and proposed alternative evaluation approaches. Trotzek et al. [41] introduced $ERDE_o^{\%}$, a variant of ERDE that normalises the evaluation based on the percentage of user writings seen before the alert. While this approach addresses the issue of user contribution normalisation, it relies on knowledge of the total number of user writings, which may not be available in real-life applications. Another proposed alternative evaluation metric for early risk prediction is $F_{latency}$, proposed by Sadeque et al. [37]. This measure aligns well with our objectives. Additionally we also used $latency_{TP}$, the delay in emitting a decision computed for the true positives. More details about the metrics can be found in [30].

3.2 Ranking-Based Evaluation

In addition to the evaluation discussed above, we employed an alternative form of evaluation to further assess the systems. After each data release (new user writing, that is post or comment), participants were required to provide the following information for each user in the collection:

- A decision for the user (alert or no alert), which was used to calculate the decision-based metrics discussed previously.
- A score representing the user's level of risk, estimated based on the evidence observed thus far.

The scores were used to create a ranking of users in descending order of estimated risk. For each participating system, a ranking was generated at each data release point, simulating a continuous re-ranking approach based on the observed evidence. In a real-life scenario, this ranking would be presented to an expert user who could make decisions based on the rankings (e.g., by inspecting the top of the rankings).

Each ranking can be evaluated using standard ranking metrics such as P@10 or NDCG. Therefore, we report the performance of the systems based on the rankings after observing different numbers of writings.

3.3 Results

Table 7 shows the participating teams, the number of runs submitted and the approximate lapse of time from the first response to the last response. This time-lapse is indicative of the degree of automation of each team's algorithms. Many of the submitted runs processed the entire thread of messages (2001), but a few variants stopped earlier. Five teams processed the thread of messages reasonably fast (less than a day for processing the entire history of user messages). The rest of the teams took several days to run the whole process.

Table 7. Task 2 (anorexia): participating teams, number of runs, number of user writings processed by the team, and lapse of time taken for the entire process.

team	#runs	#user writings processed	lapse of time (from 1st to last response)
BioNLP-IISERB [38]	5	10	09:39
GVIS	5	352	3 days 12:36
Riewe-Perla [36]	5	2001	2 days 11:25
UNSL [40]	3	2001	07:00
UMU [25]	5	2001	06:34
COS-470-Team-2	5	1	–
ELiRF-UPV [39]	4	2001	12:27
NLP-UNED [6]	5	2001	09:40
SINAI [22]	5	2001	3 days 23:49
APB-UC3M [4]	2	2001	6 days 21:34

Table 8 reports the decision-based performance achieved by the participating teams. In terms of $F1$ and latency-weighted $F1$, the best performing team was NLP-UNED (run 1), while Riewe-Perla was the team that submitted the best run (run 0) in terms of the ERDE metrics. The majority of teams made quick decisions. Overall, these findings indicate that some systems achieved a relatively high level of effectiveness with only a few user submissions. Social and public health systems may use the best predictive algorithms to assist expert humans in detecting signs of anorexia as early as possible.

Table 9 presents the ranking-based results. UNSL (run 1) obtained the best overall values after only one writing, while NLP-UNED (run 3) obtained the highest scores after 100 writings. These two teams contributed also with the best performing variants for the 500 and 1000 cutoffs.

4 Task 3: Measuring the Severity of Eating Disorders

The objective of the task is to estimate the severity of various symptoms related to the diagnosis of eating disorders. Participants were provided with a thread of user submissions to work with. For each user, a history of posts and comments from Social Media was given, and participants had to estimate the user's responses to a standardised eating disorder questionnaire based on the evidence found in the history of posts/comments.

The questionnaire used in the task is derived from the Eating Disorder Examination Questionnaire (EDE-Q)[3], which is a self-reported questionnaire consisting of 28 items. It is adapted from the semi-structured interview Eating Disorder Examination (EDE)[4] [7]. For this task, we focused on questions 1-12 and 19-28 from the EDE-Q. This questionnaire is designed to assess various aspects and severity of features associated with eating disorders. It includes four subscales: Restraint, Eating Concern, Shape Concern, and Weight Concern, along with a global score. Table 10 shows a excerpt of the EDE-Q.

The primary objective of this task was to explore the possibility of automatically estimating the severity of multiple symptoms related to eating disorders. The algorithms are required to estimate the user's response to each individual question based on their writing history. To evaluate the performance of the participating systems, we collected questionnaires completed by Social Media users along with their corresponding writing history. The user-completed questionnaires serve as the ground truth against which the responses provided by the systems are evaluated.

During the training phase, participants were provided with data from 28 users from the 2022 edition and 46 users from the 2023 edition. This training data included the writing history of the users as well as their responses to the EDE-Q questions. In the test phase, there were 18 new users for whom the participating systems had to generate results. The results were expected to follow the following specific file structure:

```
username1 answer1 answer2...answer12 answer19...answer28
username2 answer1 answer2...answer12 answer19...answer28
⋮
```

[3] https://www.corc.uk.net/media/1273/ede-q_quesionnaire.pdf.
[4] https://www.corc.uk.net/media/1951/ede_170d.pdf.

Table 8. Decision-based evaluation for Task 2

Team	Run	P	R	$F1$	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	$speed$	$latency$-$weighted$ $F1$
BioNLP-IISERB	0	0.53	0.23	0.32	0.10	0.09	2.00	**1.00**	0.32
BioNLP-IISERB	1	0.54	0.75	0.62	0.08	0.04	4.00	0.99	0.62
BioNLP-IISERB	2	0.58	0.16	0.25	0.10	0.10	**1.00**	**1.00**	0.25
BioNLP-IISERB	3	0.67	0.51	0.58	0.08	0.06	3.00	0.99	0.58
BioNLP-IISERB	4	**0.73**	0.62	0.67	0.08	0.05	4.00	0.99	0.66
GVIS	0	0.12	**1.00**	0.21	0.12	0.10	**1.00**	**1.00**	0.21
GVIS	1	0.12	**1.00**	0.22	0.12	0.10	**1.00**	**1.00**	0.22
GVIS	2	0.12	**1.00**	0.22	0.12	0.10	**1.00**	**1.00**	0.22
GVIS	3	0.12	**1.00**	0.22	0.12	0.10	**1.00**	**1.00**	0.22
GVIS	4	0.12	**1.00**	0.22	0.12	0.10	**1.00**	**1.00**	0.22
Riewe-Perla	0	0.45	0.97	0.62	**0.07**	**0.02**	6.00	0.98	0.60
Riewe-Perla	1	0.47	0.95	0.63	0.10	0.03	6.00	0.98	0.62
Riewe-Perla	2	0.47	0.95	0.63	0.10	0.03	6.00	0.98	0.62
Riewe-Perla	3	0.47	0.95	0.63	0.10	0.03	6.00	0.98	0.62
Riewe-Perla	4	0.47	0.95	0.63	0.10	0.03	6.00	0.98	0.62
UNSL	0	0.35	0.99	0.52	0.14	0.03	12.00	0.96	0.49
UNSL	1	0.42	0.96	0.59	0.14	0.03	12.00	0.96	0.56
UNSL	2	0.42	0.97	0.59	0.14	0.03	12.00	0.96	0.56
UMU	0	0.14	0.99	0.25	0.20	0.09	18.00	0.93	0.23
UMU	1	0.15	0.99	0.26	0.19	0.09	27.00	0.90	0.24
UMU	2	0.14	0.99	0.25	0.20	0.09	19.00	0.93	0.23
UMU	3	0.15	0.99	0.27	0.19	0.09	28.00	0.90	0.24
UMU	4	0.16	0.98	0.27	0.19	0.10	35.50	0.87	0.23
COS-470-Team-2	0	0.00	0.00	0.00	0.12	0.12			
COS-470-Team-2	1	0.00	0.00	0.00	0.12	0.12			
COS-470-Team-2	2	0.00	0.00	0.00	0.12	0.12			
COS-470-Team-2	3	0.00	0.00	0.00	0.12	0.12			
COS-470-Team-2	4	0.00	0.00	0.00	0.12	0.12			
ELiRF-UPV	0	0.43	0.99	0.60	0.10	0.04	12.00	0.96	0.57
ELiRF-UPV	1	0.41	**1.00**	0.58	0.10	0.04	12.00	0.96	0.56
ELiRF-UPV	2	0.32	0.99	0.49	0.12	0.04	10.00	0.96	0.47
ELiRF-UPV	3	0.43	0.99	0.60	0.11	0.04	15.00	0.94	0.57
NLP-UNED	0	0.64	0.97	0.77	0.09	0.04	13.00	0.95	0.73
NLP-UNED	1	0.67	0.97	**0.79**	0.09	0.04	14.00	0.95	**0.75**
NLP-UNED	2	0.63	0.97	0.76	0.09	0.04	12.00	0.96	0.73
NLP-UNED	3	0.63	0.98	0.77	0.09	0.03	11.00	0.96	0.74
NLP-UNED	4	0.63	0.97	0.76	0.09	0.04	14.00	0.95	0.72
SINAI	0	0.21	0.92	0.34	0.10	0.07	3.00	0.99	0.34
SINAI	1	0.21	0.92	0.34	0.10	0.07	3.00	0.99	0.34
SINAI	2	0.21	0.92	0.34	0.10	0.07	3.00	0.99	0.34
SINAI	3	0.12	**1.00**	0.21	0.13	0.10	2.00	**1.00**	0.21
SINAI	4	0.12	**1.00**	0.21	0.13	0.10	2.00	**1.00**	0.21
APB-UC3M	0	0.17	0.99	0.28	0.15	0.08	9.00	0.97	0.28
APB-UC3M	1	0.15	0.99	0.26	0.13	0.09	2.00	**1.00**	0.26

Each line has the username and 22 values (no answers from 13 to 18). These values correspond with the responses to the questions above (the possible values are 0, 1, 2, 3, 4, 5, 6).

Table 9. Ranking-based evaluation for Task 2

Team	Run	1 writing			100 writings			500 writings			1000 writings		
		$P@10$	$NDCG@10$	$NDCG@100$	$P@10$	$NDCG@10$	$NDCG@100$	$P@10$	$NDCG@10$	$NDCG@100$	$P@10$	$NDCG@10$	$NDCG@100$
BioNLP-IISERB	0	0.10	0.19	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BioNLP-IISERB	1	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BioNLP-IISERB	2	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BioNLP-IISERB	3	0.10	0.06	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BioNLP-IISERB	4	0.20	0.21	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GVIS	0	0.40	0.37	0.40	0.20	0.18	0.23	0.00	0.00	0.00	0.00	0.00	0.00
GVIS	1	0.40	0.37	0.40	0.30	0.32	0.42	0.00	0.00	0.00	0.00	0.00	0.00
GVIS	2	0.40	0.37	0.40	0.30	0.32	0.42	0.00	0.00	0.00	0.00	0.00	0.00
GVIS	3	0.40	0.37	0.40	0.30	0.32	0.42	0.00	0.00	0.00	0.00	0.00	0.00
GVIS	4	0.40	0.37	0.40	0.30	0.32	0.42	0.00	0.00	0.00	0.00	0.00	0.00
Riewe-Perla	0	0.50	0.47	0.17	0.70	0.62	0.74	0.70	0.62	0.74	0.70	0.62	0.75
Riewe-Perla	1	0.50	0.47	0.17	0.70	0.62	0.74	0.70	0.62	0.74	0.70	0.62	0.75
Riewe-Perla	2	0.50	0.47	0.17	0.70	0.62	0.74	0.70	0.62	0.74	0.70	0.62	0.75
Riewe-Perla	3	0.50	0.47	0.17	0.70	0.62	0.74	0.70	0.62	0.74	0.70	0.62	0.75
Riewe-Perla	4	0.50	0.47	0.17	0.70	0.62	0.74	0.70	0.62	0.74	0.70	0.62	0.75
UNSL	0	0.90	0.81	0.63	**1.00**	**1.00**	0.81	**1.00**	**1.00**	0.77	**1.00**	**1.00**	0.76
UNSL	1	**1.00**	**1.00**	**0.69**	**1.00**	**1.00**	0.80	0.90	0.81	0.69	0.80	0.88	0.72
UNSL	2	0.40	0.38	0.42	0.90	0.92	0.71	0.80	0.85	0.69	0.80	0.84	0.68
UMU	0	0.20	0.12	0.14	0.10	0.06	0.03	0.00	0.00	0.05	0.20	0.21	0.12
UMU	1	0.20	0.12	0.14	0.10	0.06	0.03	0.00	0.00	0.05	0.20	0.21	0.12
UMU	2	0.20	0.12	0.14	0.00	0.00	0.02	0.00	0.00	0.06	0.00	0.00	0.06
UMU	3	0.20	0.12	0.14	0.00	0.00	0.02	0.00	0.00	0.06	0.00	0.00	0.06
UMU	4	0.20	0.12	0.14	0.00	0.00	0.02	0.00	0.00	0.06	0.00	0.00	0.06
COS-470-Team-2	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
COS-470-Team-2	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
COS-470-Team-2	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
COS-470-Team-2	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
COS-470-Team-2	4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ELiRF-UPV	0	0.20	0.12	0.14	0.20	0.13	0.14	0.20	0.13	0.14	0.20	0.13	0.14
ELiRF-UPV	1	0.10	0.19	0.17	0.20	0.14	0.15	0.20	0.25	0.14	0.10	0.19	0.11
ELiRF-UPV	2	0.10	0.07	0.13	0.20	0.14	0.15	0.20	0.25	0.14	0.10	0.06	0.10
ELiRF-UPV	3	0.00	0.00	0.11	0.20	0.14	0.15	0.20	0.25	0.14	0.10	0.06	0.10
NLP-UNED	0	**1.00**	**1.00**	0.44	**1.00**	**1.00**	0.89	**1.00**	1.00	0.91	**1.00**	**1.00**	0.91
NLP-UNED	1	**1.00**	**1.00**	0.44	**1.00**	**1.00**	0.89	**1.00**	1.00	**0.92**	**1.00**	**1.00**	**0.92**
NLP-UNED	2	**1.00**	**1.00**	0.44	**1.00**	**1.00**	0.89	**1.00**	1.00	0.91	**1.00**	**1.00**	0.91
NLP-UNED	3	**1.00**	**1.00**	0.45	**1.00**	**1.00**	**0.91**	**1.00**	**1.00**	0.91	**1.00**	**1.00**	0.89
NLP-UNED	4	**1.00**	**1.00**	0.44	**1.00**	**1.00**	0.89	**1.00**	**1.00**	0.91	**1.00**	**1.00**	0.91
SINAI	0	0.00	0.00	0.07	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.00	0.03
SINAI	1	0.00	0.00	0.07	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.00	0.03
SINAI	2	0.00	0.00	0.07	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.00	0.03
SINAI	3	0.00	0.00	0.07	0.10	0.07	0.06	0.00	0.00	0.07	0.00	0.00	0.07
SINAI	4	0.00	0.00	0.07	0.10	0.07	0.06	0.00	0.00	0.07	0.00	0.00	0.07
APB-UC3M	0	0.00	0.00	0.03	0.40	0.56	0.26	0.00	0.00	0.09	0.00	0.00	0.13
APB-UC3M	1	0.10	0.06	0.07	0.00	0.00	0.18	0.00	0.00	0.10	0.00	0.00	0.08

4.1 Evaluation Metrics

Evaluation is based on the following effectiveness metrics:

– **Mean Zero-One Error (**$MZOE$**)** between the questionnaire filled by the real user and the questionnaire filled by the system (i.e. fraction of incorrect predictions).

$$MZOE(f, Q) = \frac{|\{q_i \in Q : R(q_i) \neq f(q_i)\}|}{|Q|} \tag{1}$$

where f denotes the classification done by an automatic system, Q is the set of questions of each questionnaire, q_i is the i-th question, $R(q_i)$ is the real user's answer for the i-th question and $f(q_i)$ is the predicted answer of the system for the i-th question. Each user produces a single $MZOE$ score and

Table 10. Eating Disorder Examination Questionarie

Instructions
The following questions are concerned with the past four weeks (28 days) only. Please read each question carefully. Please answer all the questions. Thank you
1. Have you been deliberately trying to limit the amount of food you eat to influence your shape or weight (whether or not you have succeeded) 0. NO DAYS 1. 1-5 DAYS 2. 6-12 DAYS 3. 13-15 DAYS 4. 16-22 DAYS 5. 23-27 DAYS 6. EVERY DAY
2. Have you gone for long periods of time (8 waking hours or more) without eating anything at all in order to influence your shape or weight? 0. NO DAYS 1. 1-5 DAYS 2. 6-12 DAYS 3. 13-15 DAYS 4. 16-22 DAYS 5. 23-27 DAYS 6. EVERY DAY
3. Have you tried to exclude from your diet any foods that you like in order to influence your shape or weight (whether or not you have succeeded)? 0. NO DAYS 1. 1-5 DAYS 2. 6-12 DAYS 3. 13-15 DAYS 4. 16-22 DAYS 5. 23-27 DAYS 6. EVERY DAY . . .
22. Has your weight influenced how you think about (judge) yourself as a person? 0. NOT AT ALL (0) 1. SLIGHTY (1) 2. SLIGHTY (2) 3. MODERATELY (3) 4. MODERATELY (4) 5. MARKEDLY (5) 6. MARKEDLY (6)
23. Has your shape influenced how you think about (judge) yourself as a person? 0. NOT AT ALL (0) 1. SLIGHTY (1) 2. SLIGHTY (2) 3. MODERATELY (3) 4. MODERATELY (4) 5. MARKEDLY (5) 6. MARKEDLY (6)
24. How much would it have upset you if you had been asked to weigh yourself once a week (no more, or less, often) for the next four weeks? 0. NOT AT ALL (0) 1. SLIGHTY (1) 2. SLIGHTY (2) 3. MODERATELY (3) 4. MODERATELY (4) 5. MARKEDLY (5) 6. MARKEDLY (6)

the reported $MZOE$ is the average over all $MZOE$ values (mean $MZOE$ over all users).

- **Mean Absolute Error (**MAE**)** between the questionnaire filled by the real user and the questionnaire filled by the system (i.e. average deviation of the predicted response from the true response).

$$MAE(f, Q) = \frac{\sum_{q_i \in Q} |R(q_i) - f(q_i)|}{|Q|} \tag{2}$$

Again, each user produces a single MAE score and the reported MAE is the average over all MAE values (mean MAE over all users).

- **Macroaveraged Mean Absolute Error (**MAE_{macro}**)** between the questionnaire filled by the real user and the questionnaire filled by the system (see [2]).

$$MAE_{macro}(f, Q) = \frac{1}{7} \sum_{j=0}^{6} \frac{\sum_{q_i \in Q_j} |R(q_i) - f(q_i)|}{|Q_j|} \tag{3}$$

where Q_j represents the set of questions whose true answer is j (note that j goes from 0 to 6 because those are the possible answers to each question). Again, each user produces a single MAE_{macro} score and the reported MAE_{macro} is the average over all MAE_{macro} values (mean MAE_{macro} over all users).

The following measures are based on aggregated scores obtained from the questionnaires. Further details about the EDE-Q instruments can be found elsewhere (e.g. see the scoring section of the questionnaire).

- **Restraint Subscale (RS):** Given a questionnaire, its restraint score is obtained as the mean response to the first five questions. This measure computes the RMSE between the restraint ED score obtained from the questionnaire filled by the real user and the restraint ED score obtained from the questionnaire filled by the system.

Each user u_i is associated with a real subscale ED score (referred to as $R_{RS}(u_i)$) and an estimated subscale ED score (referred to as $f_{RS}(u_i)$). This metric computes the RMSE between the real and an estimated subscale ED scores as follows:

$$RMSE(f, U) = \sqrt{\frac{\sum_{u_i \in U} (R_{RS}(u_i) - f_{RS}(u_i))^2}{|U|}} \tag{4}$$

where U is the user set.

- **Eating Concern Subscale (ECS):** Given a questionnaire, its eating concern score is obtained as the mean response to the following questions (7, 9, 19, 21, 20). This metric computes the RMSE (Eq. 5) between the eating concern ED score obtained from the questionnaire filled by the real user and the eating concern ED score obtained from the questionnaire filled by the system.

$$RMSE(f, U) = \sqrt{\frac{\sum_{u_i \in U}(R_{ECS}(u_i) - f_{ECS}(u_i))^2}{|U|}} \tag{5}$$

- **Shape Concern Subscale (SCS):** Given a questionnaire, its shape concern score is obtained as the mean response to the following questions (6, 8, 23, 10, 26, 27, 28, 11). This metric computes the RMSE (Eq. 6) between the shape concern ED score obtained from the questionnaire filled by the real user and the shape concern ED score obtained from the questionnaire filled by the system.

$$RMSE(f, U) = \sqrt{\frac{\sum_{u_i \in U}(R_{SCS}(u_i) - f_{SCS}(u_i))^2}{|U|}} \tag{6}$$

- **Weight Concern Subscale (WCS):** Given a questionnaire, its weight concern score is obtained as the mean response to the following questions (22, 24, 8, 25, 12). This metric computes the RMSE (Eq. 7) between the weight concern ED score obtained from the questionnaire filled by the real user and the weight concern ED score obtained from the questionnaire filled by the system.

$$RMSE(f, U) = \sqrt{\frac{\sum_{u_i \in U}(R_{WCS}(u_i) - f_{WCS}(u_i))^2}{|U|}} \tag{7}$$

- **Global ED (GED):** To obtain an overall or 'global' score, the four subscales scores are summed and the resulting total divided by the number of subscales (i.e. four) [7]. This metric computes the RMSE between the real and an estimated global ED scores as follows:

$$RMSE(f, U) = \sqrt{\frac{\sum_{u_i \in U}(R_{GED}(u_i) - f_{GED}(u_i))^2}{|U|}} \tag{8}$$

4.2 Results

Table 11. Task 3 Results. Participating teams and runs with corresponding scores for the metrics.

team	run ID	MAE	MZOE	MAE_{macro}	GED	RS	ECS	SCS	WCS
baseline	all 0 s	3.790	0.813	4.254	4.472	3.869	4.479	4.363	3.361
baseline	all 6 s	1.937	0.551	3.018	3.076	3.352	2.868	3.029	2.472
baseline	average	1.965	0.884	1.973	2.337	2.486	1.559	2.002	1.783
APB-UC3M [4]	0	2.003	0.869	2.142	2.647	2.253	1.884	2.101	1.823
DSGT [8]	0	1.965	0.588	1.713	2.211	2.321	1.969	1.944	2.117
RELAI [21]	0	2.331	0.914	2.243	2.394	2.222	2.324	2.340	**1.812**
RELAI	1	2.346	0.917	2.237	2.507	2.199	2.216	2.328	1.836
RELAI	2	2.758	0.934	2.885	2.883	2.767	3.126	3.061	2.171
RELAI	3	2.356	0.775	2.700	2.928	3.266	2.106	2.821	2.310
RELAI	4	2.851	0.884	2.979	3.159	2.784	3.150	3.068	2.336
SCaLAR-NITK [35]	0	1.912	0.591	**1.643**	2.495	2.713	1.568	**1.536**	2.098
SCaLAR-NITK	1	1.980	0.664	1.972	2.570	2.562	**1.553**	1.960	2.066
SCaLAR-NITK	2	1.879	**0.568**	1.942	2.158	2.477	2.222	2.245	2.364
SCaLAR-NITK	3	1.932	0.586	1.868	**2.117**	2.430	2.046	2.242	2.407
SCaLAR-NITK	4	**1.874**	0.672	1.820	2.292	**2.140**	1.557	1.880	2.061
UMU [25]	0	2.366	0.798	2.833	3.261	3.285	2.659	2.771	2.218
UMU	1	2.227	0.859	2.286	2.326	2.911	2.142	2.560	2.026

Table 11 reports the results obtained by the participants in this task. In order to provide some context, the table includes the performance of three baseline variants in the top block: “all 0 s”, “all 6 s”, and “average”. The “all 0 s” variant represents a strategy where the same response (0) is submitted for all questions. Similarly, the “all 6 s” variant submits the response 6 for all questions. The “average” variant calculates the mean of the responses provided by all participants for each question and submits the response that is closest to this mean value (e.g. if the mean response provided by the participants equals 3.7 then this average approach would submit a 4).

The results indicate that the top-performing system in terms of Mean Absolute Error (MAE) was run 4 by SCaLAR-NITK. This team also got the best MZOE (run 2), the best MAE_{macro} (run 0), the best GED (run 3), the best RS (run 4), the best ECS (run 1), and the best SCS (run 0). The best WCS, instead, was achieved by team RELAI (run 0). In some cases the best participating system was not better that some of the baselines (e.g., lowest MZOE is the “all 6 s” baseline).

5 Conclusions

This paper provided an overview of eRisk 2024, the eigth edition of the lab, which focused on three types of tasks: symptoms search (Task 1 on depression), early detection (Task 2 on anorexia), and severity estimations (Task 3 on eating disorders). Participants in Task 1 were given a collection of sentences and had to rank them according to their relevance to each one of the BDI-II depression symptoms. Participants in Task 2 had sequential access to social media posts and had to send alerts about individuals showing risks of anorexia. In Task 3, participants were given the full user history and had to automatically estimate the user's responses to a standard depression questionnaire.

A total of 87 runs were submitted by 17 teams for the proposed tasks. The experimental results demonstrate the value of extracting evidence from social media, indicating that automatic or semi-automatic screening tools to detect at-risk individuals could be promising. These findings highlight the need for the development of benchmarks for text-based risk indicator screening.

Acknowledgements. This work was supported by project PLEC2021-007662 (MCIN/AEI/10.13039/ 501100011033, Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Plan de Recuperación, Transformación y Resiliencia, Unión Europea-Next Generation EU). The first and second authors thank the financial support supplied by the Xunta de Galicia-Consellería de Cultura, Educación, Formación Profesional e Universidade (GPC ED431B 2022/33) and the European Regional Development Fund and project PID2022-137061OB-C21 (MCIN/AEI/10.13039/501100011033, Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Proyectos de Generación de Conocimiento; supported by "ERDF A way of making Europe", by the "European Union"). The CITIC, as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, receives subsidies from the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia. Additionally, CITIC is co-financed by the EU through the FEDER Galicia 2021–27 operational program (Ref. ED431G 2023/01). The third author thanks the financial support supplied by the Xunta de Galicia-Consellería de Cultura, Educación, Formación Profesional e Universidade (accreditation 2019–2022 ED431G-2019/04, ED431C 2022/19) and the European Regional Development Fund, which acknowledges the CiTIUS-Research Center in Intelligent Technologies of the University of Santiago de Compostela as a Research Center of the Galician University System. David E. Losada also thanks the financial support obtained from project SUBV23/00002 (Ministerio de Consumo, Subdirección General de Regulación del Juego) and project PID2022-137061OB-C22 (Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Proyectos de Generación de Conocimiento; supported by the European Regional Development Fund).

References

1. Ang, B.H., Gollapalli, S.D., Ng, S.K.: NUS-IDS@eRisk2024: ranking sentences for depression symptoms using early maladaptive schemas and ensembles. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France (2024)
2. Baccianella, S., Esuli, A., Sebastiani, F.: Evaluation measures for ordinal regression, pp. 283–287 (2009). https://doi.org/10.1109/ISDA.2009.230

3. Barachanou, A., Tsalakanidou, F., Papadopoulos, S.: REBECCA at eRisk 2024: search for symptoms of depression using sentence embeddings and prompt-based filtering. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France (2024)
4. Bascuñana, A.P., Bedmar, I.S.: APB-UC3M at eRisk 2024: natural language processing and deep learning for the early detection of mental disorders. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France (2024)
5. Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., Erbaugh, J.: An inventory for measuring depression. JAMA Psychiatry **4**(6), 561–571 (1961)
6. Fabregat, H., Deniz, D., Duque, A., Araujo, L., Martinez-Romo, J.: NLP-UNED at eRisk 2024: approximate nearest neighbors with encoding refinement for early detecting signs of anorexia. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France (2024)
7. Fairburn, C.G., Cooper, Z., O'Connor, M.: Eating disorder examination Edition 17.0D (2014)
8. Guecha, D., Potdar, A., Miyaguchi, A.: DS@GT eRisk 2024 working notes. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France (2024)
9. Hanciu, R.M.: MindwaveML at eRisk 2024: identifying depression symptoms in reddit users. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France (2024)
10. Losada, D.E., Crestani, F.: A test collection for research on depression and language use. In: Proceedings Conference and Labs of the Evaluation Forum CLEF 2016, Evora, Portugal (2016)
11. Losada, D.E., Crestani, F., Parapar, J.: eRISK 2017: CLEF lab on early risk prediction on the internet: experimental foundations. In: Jones, G.J.F., et al. (eds.) CLEF 2017. LNCS, vol. 10456, pp. 346–360. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-65813-1_30
12. Losada, D.E., Crestani, F., Parapar, J.: eRISK 2017: CLEF lab on early risk prediction on the internet: experimental foundations. In: CEUR Proceedings of the Conference and Labs of the Evaluation Forum, CLEF 2017. Dublin, Ireland (2017)
13. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2018: early risk prediction on the internet (extended lab overview). In: CEUR Proceedings of the Conference and Labs of the Evaluation Forum, CLEF 2018, Avignon, France (2018)
14. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk: early risk prediction on the internet. In: Bellot, P., et al. (eds.) CLEF 2018. LNCS, vol. 11018, pp. 343–361. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98932-7_30
15. Losada, D.E., Crestani, F., Parapar, J.: Early detection of risks on the internet: an exploratory campaign. In: Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (eds.) ECIR 2019, Part II. LNCS, vol. 11438, pp. 259–266. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-15719-7_35
16. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2019 early risk prediction on the internet. In: Crestani, F., et al. (eds.) CLEF 2019. LNCS, vol. 11696, pp. 340–357. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28577-7_27
17. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk at CLEF 2019: early risk prediction on the Internet (extended overview). In: CEUR Proceedings of the Conference and Labs of the Evaluation Forum, CLEF 2019, Lugano, Switzerland (2019)

18. Losada, D.E., Crestani, F., Parapar, J.: eRisk 2020: self-harm and depression challenges. In: Jose, J.M., et al. (eds.) ECIR 2020. LNCS, vol. 12036, pp. 557–563. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45442-5_72
19. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2020: early risk prediction on the internet. In: Arampatzis, A., et al. (eds.) CLEF 2020. LNCS, vol. 12260, pp. 272–287. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58219-7_20
20. Losada, D.E., Crestani, F., Parapar, J.: Overview of erisk at CLEF 2020: early risk prediction on the internet (extended overview). In: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece (2020)
21. Maupomé, D., Ferstler, Y., Mosser, S., Meurs, M.J.: Automatically finding evidence, predicting answers in mental health self-report questionnaires. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France (2024)
22. Mármol-Romero, A.M., Adrián Moreno-Muñoz, P.A.O., Valencia-Segura, K.M., Martínez-Cámara, E., García-Vega, M., Montejo-Ráez, A.: SINAI at eRisk@ CLEF 2024: Approaching the Search for Symptoms of Depression and Early Detection of Anorexia Signs using Natural Language Processing. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France (2024)
23. Otero, D., Parapar, J., Barreiro, Á.: Beaver: efficiently building test collections for novel tasks. In: Proceedings of the First Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2020), Samatan, Gers, France (2020)
24. Otero, D., Parapar, J., Barreiro, Á.: The wisdom of the rankers: a cost-effective method for building pooled test collections without participant systems. In: SAC 2021: The 36th ACM/SIGAPP Symposium on Applied Computing, Virtual Event, Republic of Korea, 22–26 March 2021, pp. 672–680 (2021)
25. Pan, R., Díaz, J.A.G., Beltrán, T.B., Valencia-Garcia, R.: UMUTeam at eRisk@CLEF 2024: fine-tuning transformer models with sentiment features for early detection and severity measurement of eating disorders. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France (2024)
26. Parapar, J., Martín-Rodilla, P., Losada, D.E., Crestani, F.: eRisk 2021: pathological gambling, self-harm and depression challenges. In: Hiemstra, D., Moens, M.-F., Mothe, J., Perego, R., Potthast, M., Sebastiani, F. (eds.) ECIR 2021, Part II. LNCS, vol. 12657, pp. 650–656. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-72240-1_76
27. Parapar, J., Martín-Rodilla, P., Losada, D.E., Crestani, F.: Overview of eRisk 2021: early risk prediction on the internet. In: Candan, K.S., et al. (eds.) CLEF 2021. LNCS, vol. 12880, pp. 324–344. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-85251-1_22
28. Parapar, J., Martín-Rodilla, P., Losada, D.E., Crestani, F.: Overview of erisk at CLEF 2021: early risk prediction on the internet (extended overview). In: Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, 21st–24th September 2021, pp. 864–887 (2021)
29. Parapar, J., Martín-Rodilla, P., Losada, D.E., Crestani, F.: eRisk 2022: pathological gambling, depression, and eating disorder challenges. In: Hagen, M., et al. (eds.) ECIR 2022. LNCS, vol. 13186, pp. 436–442. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-99739-7_54

30. Parapar, J., Martín-Rodilla, P., Losada, D.E., Crestani, F.: Overview of erisk 2022: early risk prediction on the internet. In: Barrón-Cedeño, A., et al. (eds.) CLEF 2022. LNCS, vol. 13390, pp. 233–256. Springer, CHam (2022). https://doi.org/10.1007/978-3-031-13643-6_18
31. Parapar, J., Martín-Rodilla, P., Losada, D.E., Crestani, F.: Overview of erisk at CLEF 2022: early risk prediction on the internet (extended overview). In: Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, 5–8 September 2022, pp. 821–850 (2022)
32. Parapar, J., Martín-Rodilla, P., Losada, D.E., Crestani, F.: erisk 2023: depression, pathological gambling, and eating disorder challenges. In: Kamps, J., et al. (eds.) ECIR 2023. LNCS, vol. 13982, pp. 585–592. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-28241-6_67
33. Parapar, J., Martín-Rodilla, P., Losada, D.E., Crestani, F.: Overview of erisk 2023: early risk prediction on the internet. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, 18–21 September 2023, pp. 233–256 (2023)
34. Parapar, J., Martín-Rodilla, P., Losada, D.E., Crestani, F.: Overview of erisk at CLEF 2023: Early risk prediction on the internet (extended overview). In: Proceedings of the Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece (2023)
35. Prasanna, S., Gulati, A.S., Karmakar, S., Hiranmayi, M.Y., Madasamy, A.K.: Measuring the severity of the signs of eating disorders using machine learning techniques. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France (2024)
36. Riewe-Perła, O., Filipowska, A.: Combining Recommender Systems and Language Models in Early Detection of Signs of Anorexia. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France (2024)
37. Sadeque, F., Xu, D., Bethard, S.: Measuring the latency of depression detection in social media. In: WSDM, pp. 495–503. ACM (2018)
38. Sarangi, P., Kumar, S., Agrawal, S., Basu, T.: A natural language processing based framework for early detection of anorexia via sequential text processing. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France (2024)
39. Segarra, A.C., Esteve, V.A., Marco, A.M., Oliver, L.F.H.: ELiRF-VRAIN at eRisk 2024: using LongFormers for early detection of signs of anorexia. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France (2024)
40. Thompson, H., Errecalde, M.: A time-aware approach to early detection of anorexia: UNSL at eRisk 2024. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France (2024)
41. Trotzek, M., Koitka, S., Friedrich, C.: Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. IEEE Trans. Knowl. Data Eng. (2018)

Overview of EXIST 2024 — Learning with Disagreement for Sexism Identification and Characterization in Tweets and Memes

Laura Plaza[1(✉)], Jorge Carrillo-de-Albornoz[1], Víctor Ruiz[1], Alba Maeso[2], Berta Chulvi[2], Paolo Rosso[2,3], Enrique Amigó[1], Julio Gonzalo[1], Roser Morante[1], and Damiano Spina[4]

[1] Universidad Nacional de Educación a Distancia (UNED), 28040 Madrid, Spain
{lplaza,jcalbornoz,victor.ruiz,enrique,julio,rmorant}@lsi.uned.es

[2] Universitat Politècnica de València (UPV), 46022 Valencia, Spain
prosso@dsic.upv.es, berta.chulvi@upv.es, amaeolm@inf.upv.es

[3] ValgrAI - Valencian Graduate School and Research Network of Artificial Intelligence, 46022 Valencia, Spain

[4] RMIT University, 3000 Melbourne, Australia
damiano.spina@rmit.edu.au

Abstract. In recent years, the rapid increase in the dissemination of offensive and discriminatory material aimed at women through social media platforms has emerged as a significant concern. This trend has had adverse effects on women's well-being and their ability to freely express themselves. The EXIST campaign has been promoting research in online sexism detection and categorization in English and Spanish since 2021. The fourth edition of EXIST, hosted at the CLEF 2024 conference, consists of three groups of tasks, which are a continuation of EXIST 2023: *sexism identification*, *source intention identification*, and *sexism categorization*. However, while EXIST 2023 focused on processing tweets, the novelty of this edition is that the three tasks are also applied to memes, resulting in a total of six tasks. The "learning with disagreement" paradigm is adopted to address disagreements in the labelling process and promote the development of equitable systems that are able to learn from different perspectives on the sexism phenomena. The 2024 edition of EXIST has exceeded the success of previous editions, with the participation of 57 teams submitting 412 runs. This lab overview describes the tasks, dataset, evaluation methodology, participant approaches and results.

Keywords: sexism identification · sexism categorization · learning with disagreement · memes · data bias

1 Introduction

EXIST (sEXism Identification in Social neTworks) is a series of scientific events and shared tasks on sexism identification in social networks. The editions of

L. Goeuriot et al. (Eds.): CLEF 2024, LNCS 14959, pp. 93–117, 2024.
https://doi.org/10.1007/978-3-031-71908-0_5

2021 and 2022 [36,37], celebrated under the umbrella of the IBERLEF forum, were the first in proposing tasks focusing on identifying and classifying online sexism in a broad sense, from explicit and/or hostile to other subtle or even benevolent expressions. The 2023 edition [33] took place as a CLEF Lab and added a third task consisting in determining the intention of the author of sexist messages with the aim of raising awareness against sexism. Additionally, the main novelty of the 2023 edition was the adoption of the "Learning with Disagreements" (LwD) paradigm [47] for the development of the dataset and for the evaluation of the systems. In the LwD paradigm, models are trained to handle and learn from conflicting or diverse annotations so that different annotators' perspectives, biases, or interpretations are taken into account. This approach fits the findings of our previous work that showed that the perception of sexism is strongly dependent on the demographic and cultural background of the individual. Adopting this paradigm was a distinguishing feature in comparison to the SemEval-2023 Shared Task 10: "Explainable Detection of Online Sexism" [18].

EXIST 2024,[1] organised also as a CLEF Lab, aims to continue contributing datasets and tasks that help developing applications to combat sexism on-line, as a form of hate on-line. This edition embraces also the LwD paradigm and, as novelty, incorporates three new tasks that center around memes. Memes are images that are spread rapidly by social networks and Internet users. While by nature memes are humorous, there is a growing tendency to use them for harmful purposes, as an strategy to conceal hate speech by combining stylistic devices of humour [5], since people tolerate humorously communicated prejudices better than explicit irrespectful remarks [8,19,24]. Thus, memes contribute to spreading derogatory humour and to strengthen preexisting prejudices and maintaining hierarchies between social groups [14]. As Gasparini et al. indicate [12], misogyny and sexism against women are widespread attitudes within the social media communities, reinforcing age-old patriarchal establishments of baseless name-calling, objectifying their appearances, and stereotyping gender roles. By including sexist memes in the EXIST 2024 dataset, we aim to encompass a broader spectrum of sexist manifestations in social networks and to contribute to the development of automated multimodal tools capable of detecting harmful content targeting women.

Meme detection has also been the focus of other competitions. The SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification [11] focused on the detection of misogynous memes on the web in English and proposed two tasks: recognising whether a meme is misogynous or not and recognising types of misogyny in memes. The Overview of Shared Task on Multitask Meme Classification - Unraveling Misogynistic and Trolls in Online Memes [7] consisted in classifying misogynistic content and troll memes, focusing specifically on memes in Tamil and Malayalam languages. The originality of EXIST lies in that the languages addressed are English and Spanish, it introduces also the task on source intention recognition and it adopts the LwD paradigm.

[1] http://nlp.uned.es/exist2024/. Accessed 28 May 2024.

In the following sections, we provide comprehensive information about the tasks, the dataset, the evaluation methodology, the results and the different approaches of the systems that participated in the EXIST 2024 Lab. The competition features six distinct tasks: sexism identification, source intention classification, and sexism categorization both in tweets and in memes. A total of 148 teams from 32 different countries registered to participate. Ultimately, we received 412 results from 57 teams.

2 Tasks

The 2024 edition of EXIST feature 6 tasks, which are described below. The languages addressed are English and Spanish and the datasets are collections of tweets and memes. For the tasks on memes, all the partitions of the dataset are new, whereas for the tasks on tweets we employ the EXIST 2023 dataset.

2.1 Task 1: Sexism Identification in Tweets

This is a binary classification task where systems must decide whether or not a given tweet expresses ideas related to sexism in any of the three forms: it is sexist itself, it describes a sexist situation in which discrimination towards women occurs, or criticizes a sexist behaviour The following statements from the dataset show examples of sexist and not sexist messages, respectively.

(1) **Sexist**. *It's less of #adaywithoutwomen and more of a day without feminists, which, to be quite honest, sounds lovely.*
(2) **Not sexist**. *Just saw a woman wearing a mask outside spank her very tightly leashed dog and I gotta say I love learning absolutely everything about a stranger in a single instant.*

2.2 Task 2: Source Intention in Tweets

This task aims to categorize the message according to the intention of the author. We propose the following ternary classification of tweets:

- **Direct** sexist message. The intention is to write a message that is sexist by itself or incites to be sexist, as in:
  (3) *A woman needs love, to fill the fridge, if a man can give this to her in return for her services (housework, cooking, etc.), I don't see what else she needs.*
- **Reported** sexist message. The intention is to report and share a sexist situation suffered by a woman or women in first or third person, as in:
  (4) *I doze in the subway, I open my eyes feeling something weird: the hand of the man sat next to me on my leg #SquealOnYourPig.*
- **Judgemental** message. The intention is to condemn sexist situations or behaviours, as in:
  (5) *As usual, the woman was the one quitting her job for the family's welfare...*

2.3 Task 3: Sexism Categorization in Tweets

Many facets of a woman's life may be the focus of sexist attitudes including domestic and parenting roles, career opportunities, sexual image, and life expectations, to name a few. According to this, each sexist tweet must be assigned one or more of the following categories:

- **Ideological and inequality**. It includes messages that discredit the feminist movement. It also includes messages that reject inequality between men and women, or present men as victims of gender-based oppression.
  (6) *#Feminism is a war on men, but it's also a war on women. It's a war on female nature, a war on femininity.*
- **Stereotyping and dominance**. It includes messages that express false ideas about women that suggest they are more suitable or inappropriate for certain tasks, and somehow inferior to men.
  (7) *Most women no longer have the desire or the knowledge to develop a high quality character, even if they wanted to.*
- **Objectification**. It includes messages where women are presented as objects apart from their dignity and personal aspects. We also include messages that assume or describe certain physical qualities that women must have in order to fulfill traditional gender roles.
  (8) *No offense but I've never seen an attractive african american hooker. Not a single one.*
- **Sexual violence**. It includes messages where sexual suggestions, requests or harassment of a sexual nature (rape or sexual assault) are made.
  (9) *I wanna touch your tits..you can't imagine what I can do on your body.*
- **Misogyny and non sexual violence**. It includes expressions of hatred and violence towards women.
  (10) *Domestic abuse is never okay... Unless your wife is a bitch.*

2.4 Task 4: Sexism Identification in Memes

As in Task 1, this is a binary classification consisting on deciding whether or not a meme is sexist. Figure 1 shows examples of sexist and non sexist memes.

2.5 Task 5: Source Intention in Memes

As in Task 2, this task aims to categorize the meme according to the intention of the author. However, in this task systems should only classify memes in two classes: DIRECT or JUDGEMENTAL, as shown in Fig. 2.

2.6 Task 6: Sexism Categorization in Memes

This task aims to classify sexist memes according to the categorization provided for Task 3: (i) ideological and inequality, (ii) stereotyping and dominance, (iii) objectification, (iv) sexual violence and (v) misogyny and non-sexual violence. Figure 3 shows one meme of each category.

(a) Sexist meme (b) Non sexist meme

Fig. 1. Examples of sexist and non-sexist memes.

(a) Direct (b) Judgemental

Fig. 2. Examples of direct and judgemental memes.

3 Dataset

The EXIST 2024 dataset comprises two types of data: the tweets from the EXIST 2023 dataset and a completely new dataset of memes. Here, we briefly describe the process followed to curate the meme dataset. More details – including bias considerations, the annotation process, quality experiments, and inter-annotator agreement – can be found in the extended overview [32]. In contrast, Plaza et al. [33] provide a detailed description of the tweet dataset.

Since we adopt the LwD paradigm, we provide all labels assigned by the different annotators to allow systems to learn from conflicting and subjective information. This paradigm not only proved to improve the systems' accuracy, robustness and generalizability, but also helped to mitigate bias.

3.1 Data Sampling

We first curated a lexicon of terms and expressions leading to sexist memes. The set of seeds encompasses diverse topics and contains 250 terms, with 112 in English and 138 in Spanish.

(a) Ideological & inequality (b) Objectification

(c) Stereotyping & dominance (d) Sexual violence (e) Misogyny & non-sexual violence

Fig. 3. Examples of memes from the different sexist categories.

The terms were used as search queries on Google Images to obtain the top 100 images. Rigorous manual cleaning procedures were applied, defining memes and ensuring the removal of noise such as textless images, text-only images, ads, and duplicates. The final set consists of more than 3,000 memes per language.

Since the proportion of memes per term was heterogeneous, we discarded the most unbalanced seeds and made sure that all seeds have at least five memes. To avoid introducing selection bias, we randomly selected memes, ensuring the appropriate distribution per seed. As a result, we have 2,000 memes per language for the training set and 500 memes per language for the test set.

3.2 Labeling with Disagreements

We have considered some sources of "label bias", that may be introduced by the socio-demographic differences of the persons that participate in the annotation process, but also when more than one possible correct label exists or when the decision on the label is highly subjective. In order to mitigate label bias, we consider two sociodemographic parameters: gender (MALE/FEMALE) and age (18–22/23–45/+46 y.o.). Each meme was annotated by 6 annotators selected through the Prolific crowdsourcing platform.[2] Also, as a new feature in

[2] No personally identifiable information about the crowd workers was collected. Crowd workers were informed that the tweets could contain offensive information and were allowed to withdraw voluntarily at any time. Full consent was obtained.

the datasets, both of 2023 and 2024, we have requested three additional demographic characteristics of annotators: level of education, ethnicity, and country of residence.

4 Evaluation Methodology and Metrics

As in EXIST 2023, we have carried out a **"soft evaluation"** and a **"hard evaluation"**. The soft evaluation relates to the LwD paradigm and is intended to measure the ability of the model to capture disagreements, by considering the probability distribution of labels in the output as a soft label and comparing it with the probability distribution of the annotations. The hard evaluation is the standard paradigm and assumes that a single label is provided by the systems for every instance in the dataset.

1. **Soft-soft evaluation**. For systems that provide probabilities for each category, we perform a soft-soft evaluation that compares the probabilities assigned by the system with the probabilities assigned by the set of human annotators. The probabilities of the classes for each instance are calculated according to the distribution of labels and the number of annotators for that instance. We use a modification of the original ICM metric (Information Contrast Measure [1]), ICM-Soft (see details below), as the official evaluation metric in this variant and we also provide results for the normalized version of ICM-Soft (ICM-Soft Norm).
2. **Hard-hard evaluation**. For systems that provide a hard, conventional output, we perform a hard-hard evaluation. To derive the hard labels in the ground truth from the different annotators' labels, we use a probabilistic threshold computed for each task. As a result, for Tasks 1 and 4, the class annotated by more than 3 annotators is selected; for Tasks 2 and 5, the class annotated by more than 2 annotators is selected; and for Tasks 3 ad 6 (multilabel), the classes annotated by more than 1 annotator are selected. The instances for which there is no majority class (i.e., no class receives more probability than the threshold) are removed from this evaluation scheme. The official metric for this task is the original ICM, as defined by Amigó and Delgado. We also report a normalized version of ICM (ICM Norm) and F1 ($\mathrm{F1}_{\mathrm{YES}}$). In Tasks 1 and 4, we use F1 for the positive class. In Tasks 2, 3, 5 and 6, we use the macro-average of F1 for all classes (Macro F1). Note, however, that F1 is not ideal in our experimental setting: although it can handle multilabel situations, it does not take into account the relationships between classes. In particular, a confusion between not sexist and any of the sexist subclasses, and a confusion between two of the sexist subclasses, are penalized equally.

ICM is a similarity function that generalizes Pointwise Mutual Information (PMI), and can be used to evaluate outputs in classification problems by computing their similarity to the ground truth. The general definition of ICM is:

$$\mathrm{ICM}(A, B) = \alpha_1 IC(A) + \alpha_2 IC(B) - \beta IC(A \cup B)$$

Where $IC(A)$ is the Information Content of the instance represented by the set of features A. ICM maps into PMI when all parameters take a value of 1. The general definition of ICM by [1] is applied to cases where categories have a hierarchical structure and instances may belong to more than one category. The resulting evaluation metric is proved to be analytically superior to the alternatives in the state of the art. The definition of ICM in this context is:

$$\text{ICM}(s(d), g(d)) = 2IC(s(d)) + 2IC(g(d)) - 3IC(s(d) \cup g(d))$$

Where $IC()$ stands for Information Content, $s(d)$ is the set of categories assigned to document d by system s, and $g(d)$ the set of categories assigned to document d in the gold standard. The score for a perfect output $(s(d) = g(d))$ is the gold standard Information Content $(IC(g(d))$. The score for a zero-information system (no category assignment) is $-IC(g(d))$. We use these two boundaries for normalisation purposes, truncating to 0 the scores lower than $-IC(g(d))$.

As there is not, to the best of our knowledge, any current metric that fits hierarchical multilabel classification problems in a LwD scenario, we have defined an extension of ICM (ICM-soft) that accepts both soft system outputs and soft ground truth assignments. ICM-soft works as follows: first, we define the Information Content of a single assignment of a category c with an agreement v to a given instance as the probability of instances in the gold standard to exceed the agrement level v for the category c:

$$IC(\{\langle c, v\rangle\}) = -\log_2(P(\{d \in D : g_c(d) \geq v\})$$

In order to estimate IC, we compute the mean and deviation of the agreement levels for each class across instances, and applying the cumulative probability over the inferred normal distribution.[3]

Due to the multi-label and hierarchical nature of the classification task, for each classification instance, the gold standard, the system output and their unions $(IC(s(d))$ $IC(g(d))$ and $IC(s(d))Ug(d))$ are sets of category assignments. The union of the assignments (i.e. $s(d))Ug(d))$ is calculated as fuzzy sets, i.e. the maximum values., in order to estimate information content, we apply a recursive function similar to the one described by Amigó and Delgado [1] for assignment sets and avoid the redundant information of parent categories.

$$IC\left(\bigcup_{i=1}^{n}\{\langle c_i, v_i\rangle\}\right) = IC(\langle c_1, v_1\rangle) + IC\left(\bigcup_{i=2}^{n}\{\langle c_i, v_i\rangle\}\right) - IC\left(\bigcup_{i=2}^{n}\{\langle \text{lca}(c_1, c_i), min(v_1, v_i)\rangle\}\right) \quad (1)$$

where $\text{lca}(a, b)$ is the lowest common ancestor of categories a and b.

[3] In the case of zero variance, we must consider that the probability for values equals or below the mean is 1 (zero IC) and the probability for values above the mean must be smoothed. But this is not the case of the EXIST datasets.

Table 1. Runs submitted and teams participating on each EXIST 2024 task.

	Tweets			Memes		
	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6
# Runs	106	77	63	87	36	43
# Teams	46	38	27	41	18	22

5 Overview of Approaches

In this section, we provide a concise overview of the approaches presented at EXIST 2024. For a comprehensive description of the systems, please refer to the participant papers and the extended overview [32]. Although 148 teams from 32 different countries registered for participation, the number of participants who finally submitted results were 57, submitting 412 runs. Teams were allowed to participate in any of the six tasks and submit hard and/or soft outputs. Table 1 summarizes the participation in the different tasks and evaluation contexts.

The evaluation campaign started on March 4, 2024 with the release of the training set. The test set was made available on April 15. The participants were provided with the official evaluation script. Runs had to be submitted by May 10. Each team could submit up to three runs per task.

A wide range of approaches and strategies were used by the participants. Here we summarize the techniques and tools employed. For a comprehensive description, please refer to the extended overview [32]. Nearly all participant systems utilized large language models, both monolingual and multilingual. Most employed LLMs include BERT, DistilBERT, MarIA, MDEBERTA, RoBERTa, DeBERTa, Llama, and GPT-4. For processing memes, popular vision models were employed: CLIP, BEIT and VIT. Some teams employed ensembles of multiple models to enhance the overall performance. A couple of teams made use of knowledge integration to combine different language models with language features. Data augmentation techniques were used by several teams. Prompt Engineering was also used to adapt pre-trained models to the sexism detection task. Only two teams utilized deep learning architectures such as BiLSTM and CNN, while one team opted for traditional machine learning methods, including SVM, Random Forest, and XGBoost, among others. As in EXIST 2023 [31], Twitter-specific models where employed, such as Twitter-RoBERTa and Twitter-XML-RoBERTa.

While 174 systems took advantage of the multiple annotations available and provided soft outputs, 238 followed the traditional approach of providing only hard labels as outputs. Textual tasks received greater engagement, although participation is also high in the meme tasks. The binary classification tasks had more participants, followed by mono-label tasks, and finally, multi-label tasks, which is due to the increasing difficulty of these tasks.

For each of the six tasks, the organization also provided two different baseline runs: **EXIST2024 majority**, a non-informative baseline that classifies all instances as the majority class; and **EXIST2024 MINORITY**, a non-informative baseline that classifies all instances as the minority class. The evaluation metrics for the gold standard (**EXIST2024 gold**) are also provided, in order to set the upper bound for the ICM metrics.

6 Results

In the next subsections, we report the results of the participants and the baseline systems for each task. We only show the results obtained by the best run submitted by each participant to each task. For more detailed results, please refer to the Lab Working Notes [32].

6.1 Task 1: Sexism Identification in Tweets

Soft Evaluation. Table 2 presents the results for the soft-soft evaluation for Task 1. A total of 37 runs were submitted. Out of these, 34 runs outperformed the non-informative majority class baseline (where all instances are labeled as "NO"), and all runs surpassed the non-informative minority class baseline (where all instances are labeled as "YES"). We observed a significant discrepancy in performance, with ICM-Soft Norm scores ranging from 0.6755 to 0.0374. However, if

Table 2. Results of Task 1 in the soft-soft evaluation (the best submission from each team).

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
EXIST2024 gold	0	3.1182	1.0000	0.5472
NYCU-NLP_1 [10]	1	1.0944	0.6755	0.9088
ABCD Team_3 [34]	4	0.9291	0.6490	1.2637
CIMAT-CS-NLP_3 [46]	5	0.9285	0.6489	1.2252
BAZI_1 [3]	9	0.8179	0.6311	0.9750
Awakened_2 [30]	10	0.7196	0.6154	0.8106
Victor-UNED_1 [38]	11	0.6952	0.6115	1.0691
I2C-UHU_2 [13]	13	0.6871	0.6102	0.9184
UMUTEAM_1 [28]	15	0.6679	0.6071	0.8708
MMICI_3 [15]	21	0.4589	0.5736	2.0316
clac_1	22	0.1431	0.5230	2.9543
RMIT-IR_1 [44]	23	–0.0011	0.4998	2.7892
FraunhoferSIT_1 [9]	24	–0.0658	0.4895	0.8801
CNLP-NITS-PP_1 [49]	25	–0.2086	0.4666	1.0390
Atresa-I2C-UHU_1 [6]	27	–0.3256	0.4478	3.9518
UniLeon-UniBO_1	33	–1.1882	0.3095	1.2449
EXIST2024 majority	36	–2.3585	0.1218	4.6115
NICA_3 [26]	37	–2.8848	0.0374	1.5286
EXIST2024 minority	40	–3.0717	0.0075	5.3572

we analyze the top 5 systems, we appreciate a difference of less than 5 percentual points. Notably, the best run achieved an ICM-Soft Norm score of 68% for this binary classification task, surpassing the top performance of 64% recorded by the best EXIST 2023 participant. This suggests that new models and approaches are becoming more effective at detecting sexism in social networks. However, it also indicates that there is still room for improvement.

Hard Evaluation. Table 3 presents the results for the hard-hard evaluation. In this scenario, the annotations from the six annotators are combined into a single label using the majority vote. Out of the 67 systems submitted for this task, 66 ranked above the majority class baseline (all instances labeled as “NO”). All systems surpassed the minority class baseline (all instances labeled as “YES”). Similar to the soft-soft evaluation, the results vary considerably. If we focus on the ICM-hard normalized metric, we observe that the best run gets 0.8002 while the worse one gests only 0.2665. If we focus on the top 5 systems, we observe that they achieve comparable results.

6.2 Task 2: Source Intention in Tweets

Soft Evaluation. Table 4 presents the results for the soft-soft evaluation of Task 2. The table shows that 32 runs were submitted. Among them, 25 runs achieved better results compared to the majority class baseline (where all instances are labeled as “NO”). Furthermore, all of the submitted runs outperformed or equaled the minority class baseline (where all instances are labeled as “REPORTED”). The ICM-Soft Norm scores range from the 0.4795 of the best system (“nycu-nlp_2”) system to 0.0000 of “fmrs_2”, indicating significant variability in the effectiveness of the submitted models. It is worth mentioning that the best system outperforms the second-best by more than 8% points. Overall, performance is considerably lower compared to Task 1. This can be attributed to the hierarchical and multiclass nature of Task 2.

Hard Evaluation. Table 5 presents the hard-hard evaluation results for Task 2, assessing 43 systems against the hard gold standard. Among these, 37 runs outperform the majority class baseline (where all instances are labeled “NO”), and all systems show equal or better performance compared to the minority class baseline (where all instances are labeled “REPORTED”). Similar to the soft-soft evaluation, discrepancies between the best and the worst-performing systems are more pronounced in Task 2 than in Task 1. The top-ranking system, “ABCD Team_1,” achieved the highest ICM-Hard normalized score (0.6320). The top 5 best systems range between 0.5937 and 0.6320. The lower end of the table includes five systems which score 0 in the ICM-Hard norm metric.

Table 3. Results of Task 1 in the hard-hard evaluation (the best submission from each team).

Run	Rank	ICM-Hard	ICM-Hard Norm	$F1_{YES}$
EXIST2024 gold	0	0.9948	1.0000	1.0000
NYCU-NLP_1	1	0.5973	0.8002	0.7944
ABCD Team_1	2	0.5957	0.7994	0.7826
CIMAT-CS-NLP_2	3	0.5926	0.7978	0.7899
EquityExplorers_2 [17]	4	0.5883	0.7957	0.7775
CIMAT-GTO_3 [50]	5	0.5848	0.7939	0.7903
I2C-UHU_2	10	0.5557	0.7793	0.7733
BAZI_1	11	0.5490	0.7759	0.7755
ADITYA_3 [39]	14	0.5418	0.7723	0.7691
MMICI_3 [15]	17	0.5324	0.7676	0.7637
NICA_1	19	0.5214	0.7621	0.7642
Awakened_3	20	0.5196	0.7611	0.7652
maven_3	22	0.5015	0.7521	0.7596
Victor-UNED_3	24	0.4934	0.7480	0.7602
RMIT-IR_3	27	0.4802	0.7414	0.7548
penta-nlp_1	29	0.4779	0.7402	0.7508
fmrs_2 [48]	35	0.4398	0.7211	0.7462
clac_1	36	0.4380	0.7201	0.7376
TextMiner_2 [16]	39	0.3926	0.6973	0.7223
CAU&ITU_2 [22]	45	0.3460	0.6739	0.7024
DLRG_1	46	0.3446	0.6732	0.7085
shm2024_3 [42]	48	0.3230	0.6623	0.7044
Atresa-I2C-UHU_1	52	0.2782	0.6398	0.6899
FraunhoferSIT_1	53	0.2320	0.6166	0.6823
CNLP-NITS-PP_1	54	0.1977	0.5994	0.6762
mc-mistral_2 [43]	56	0.0614	0.5309	0.5317
UniLeon-UniBO_2	58	–0.1870	0.4060	0.4963
NIT-Patna-NLP_1	60	–0.2975	0.3505	0.5272
shm2024_2	62	–0.3410	0.3286	0.4922
DadJokers_1	63	–0.3611	0.3185	0.4365
VerbaNex_1 [23]	64	–0.4048	0.2965	0.4588
The 3 Musketeers_1 [45]	65	–0.4229	0.2875	0.3371
EXIST2024 majority	68	–0.4413	0.2782	0.0000
EXIST2024 minority	70	–0.5742	0.2114	0.5698

Table 4. Results of Task 2 in the soft-soft evaluation (the best submission from each team).

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
EXIST2024 gold	0	6.2057	1.0000	0.9128
NYCU-NLP_2	1	–0.2543	0.4795	1.8344
BAZI_1	4	–1.3468	0.3915	1.7812
Victor-UNED_2	5	–1.6440	0.3675	1.7971
ABCD Team_3	7	–1.8462	0.3513	2.4123
UMUTEAM_1	8	–1.9566	0.3424	1.4726
Awakened_2	9	–2.0091	0.3381	3.0835
fmrs_1	14	–2.1737	0.3249	2.1210
CNLP-NITS-PP_1	15	–2.4732	0.3007	1.6696
Atresa-I2C-UHU_1	16	–2.6802	0.2841	2.1629
I2C-UHU_2	17	–2.6952	0.2828	2.1440
MMICI_3	20	–3.6350	0.2071	1.7285
FraunhoferSIT_1	21	–4.0856	0.1708	1.7649
RMIT-IR_1	23	–4.5481	0.1336	3.5776
CUET-SSTM_1	26	–5.1320	0.0865	4.8736
EXIST2024 majority	27	–5.4460	0.0612	4.6233
NICA_2	28	–5.7592	0.0360	2.7026
UniLeon-UniBO_3	30	–5.7633	0.0356	2.1267
EXIST2024 minority	35	–32.9552	0.0000	8.8517

6.3 Task 3: Sexism Categorization in Tweets

Soft Evaluation. Table 6 displays the results of the soft-soft evaluation for Task 3. A total of 30 runs were submitted, with 26 runs surpassing the majority class baseline (all instances labeled as “NO”), and all systems outperforming the minority class baseline (all instances labeled as “SEXUAL-VIOLENCE”). The “NYCU-NLP” team has the top three runs, with ”NYCU-NLP_1“ ranked first (ICM-Soft: −1.1762, ICM-Soft Norm: 0.4379). The next two runs from the same team, ”NYCU-NLP_2“ and “NYCU-NLP_3,” follow closely, indicating the consistency and robustness of their approach. The fourth and fifth systems, however, show a significantly poorer performance (0.3835 and 0.3732, respectively) The range of ICM-Soft Norm scores (from 0.4379 to 0.0000) underscores a significant variability in system performance. However, despite the complexity of the task, it seems that systems are still able to correctly capture relevant information concerning the different types of sexism.

Hard Evaluation. In the hard-hard evaluation context for the third task, 31 systems were submitted. As shown in Table 7, 28 systems outperformed

Table 5. Results of Task 2 in the hard-hard evaluation (the best submission from each team).

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
EXIST2024 gold	0	1.5378	1.0000	1.0000
ABCD Team_1	1	0.4059	0.6320	0.5677
NYCU-NLP_3	2	0.3522	0.6145	0.5410
CUET-SSTM_1	5	0.2883	0.5937	0.5383
CIMAT-CS-NLP_2	7	0.2643	0.5859	0.5171
penta-nlp_1	9	0.2089	0.5679	0.4856
BAZI_1	10	0.1883	0.5612	0.4843
I2C-UHU_2	11	0.1815	0.5590	0.4980
Awakened_2	12	0.1812	0.5589	0.4826
fmrs_1	14	0.1609	0.5523	0.4978
NICA_2	15	0.1506	0.5490	0.4738
RMIT-IR_1	18	0.0855	0.5278	0.4024
Victor-UNED_1	19	0.0851	0.5277	0.3257
maven_1 [40]	26	–0.0510	0.4834	0.4563
MMICI_1	27	–0.0987	0.4679	0.4548
DLRG_1	29	–0.1171	0.4619	0.3931
Atresa-I2C-UHU_1	31	–0.1524	0.4504	0.4278
CNLP-NITS-PP_1	33	–0.2694	0.4124	0.3743
FraunhoferSIT_1	34	–0.4106	0.3665	0.3823
CAU&ITU_1	35	–0.4711	0.3468	0.2998
shm2024_1	37	–0.8873	0.2115	0.3148
EXIST2024 majority	39	–0.9504	0.1910	0.1603
UniLeon-UniBO_3	41	–1.2145	0.1051	0.2605
NIT-Patna-NLP_1	42	–1.9410	0.0000	0.1207
EXIST2024 minority	46	–3.1545	0.0000	0.0280

the majority class baseline (all instances labeled as "NO"), while all systems achieved better results than the minority class baseline (all instances labeled as "SEXUAL-VIOLENCE"). The discrepancy between the best ("ABCD Team_1", 0.5862 ICM-Hard norm score) and the worst-performing system ("CAU&ITU" 1, 0.000 score) is over 0.5 ICM-hard-norm, which is less than in Task 2. Finally, comparing the performance of the three different textual tasks in the hard-hard evaluation, the efficiency of the systems in this task, in terms of ICM-Hard Norm, is lower than in previous tasks. This further highlights the complexity of categorizing sexism.

Table 6. Results of Task 3 in the soft-soft evaluation (the best submission from each team).

Run	Rank	ICM-Soft	ICM-Soft Norm
EXIST2024 gold	0	9.4686	1.0000
NYCU-NLP_1	1	–1.1762	0.4379
Medusa_1 [2]	4	–2.2055	0.3835
ABCD Team_3	7	–3.5160	0.3143
Awakened_2	9	–4.0748	0.2848
NICA_2	12	–4.4324	0.2659
FraunhoferSIT_1	14	–5.1905	0.2259
Victor-UNED_1	15	–5.5936	0.2046
CNLP-NITS-PP_1	17	–5.7385	0.1970
RMIT-IR_1	19	–7.2098	0.1193
MMICI_3	20	–7.6413	0.0965
fmrs_1	24	–8.2508	0.0643
EXIST2024 majority	28	–8.7089	0.0401
UniLeon-UniBO_1	29	–10.3622	0.0000
Atresa-I2C-UHU_1	32	–10.4052	0.0000
EXIST2024 minority	33	–46.1080	0.0000

Table 7. Results of Task 3 in the hard-hard evaluation (the best submission from each team).

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
EXIST2024 gold	0	2.1533	1.0000	1.0000
ABCD Team_1	1	0.3713	0.5862	0.6004
NYCU-NLP_3	3	0.3069	0.5713	0.6130
Awakened_2	6	–0.0042	0.4990	0.4833
RMIT-IR_3	8	–0.0344	0.4920	0.5049
ABCD Team_2	12	–0.1090	0.4747	0.5286
NICA_2	13	–0.2383	0.4447	0.4564
penta-nlp_1 [41]	14	–0.2597	0.4397	0.4379
maven_1	15	–0.2654	0.4384	0.4491
UniLeon-UniBO_1	16	–0.3188	0.4260	0.5032
UMUTEAM_1	20	–0.7339	0.3296	0.4942
FraunhoferSIT_1	21	–0.7437	0.3273	0.3724
MMICI_3	23	–0.8105	0.3118	0.4805
CNLP-NITS-PP_1	25	–0.9571	0.2778	0.2684
fmrs_3	29	–1.5952	0.1296	0.1087
EXIST2024 majority	30	–1.5984	0.1289	0.1069
CAU&ITU_1	33	–2.3423	0.0000	0.1705
EXIST2024 minority	34	–3.1295	0.0000	0.0288

6.4 Task 4: Sexism Identification in Memes

Soft Evaluation. Table 8 presents the results for the classification of memes as sexist or not sexist. The performance results are notably low for a binary classification task: "Victor-UNED_1", the top-ranked participant, achieved an ICM-Soft Norm score of 0.4530 and a relatively low Cross Entropy of 1.1028. However, the variability between the best and worst-performing systems is reduced compared to that of the tasks described above. When comparing these results to those of Task 1 (classifying tweets as sexist or not), we observe a significant drop in performance for image classification (0.4530 versus 0.6755 ICM-Soft Norm). It is important to highlight that most approaches relied solely on the text within the meme for classification, without incorporating image processing. This suggests that sexism in memes might often be conveyed through the imagery, even when the accompanying text seems to be neutral.

Table 8. Results of Task 4 in the soft-soft evaluation (the best submission from each team).

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
EXIST2024 gold	0	3.1107	1.0000	0.5852
Victor-UNED_1	1	–0.2925	0.4530	1.1028
Elias&Sergio_1	3	–0.3225	0.4482	0.9903
I2C-Huelva_3	4	–0.3263	0.4476	1.5189
RMIT-IR_2	8	–0.3780	0.4392	0.9852
NICA_1	9	–0.4360	0.4299	0.9278
PINK_2 [35]	10	–0.4396	0.4293	0.9375
ROCurve_3	12	–0.4646	0.4253	0.9609
the gym nerds_2	13	–0.5015	0.4194	0.9201
MMICI_2	16	–0.6183	0.4006	0.9143
OppositionalOppotision_1	21	–0.9556	0.3464	3.2025
melialo-vcassan_1	22	–1.0022	0.3389	0.9931
CNLP-NITS-PP_2	27	–1.2354	0.3014	1.0918
CHEEXIST_2	28	–1.2710	0.2957	1.1993
Penta-ML_2 [4]	30	–1.2910	0.2925	2.2277
epistemologos_1	31	–1.3486	0.2832	2.9425
EXIST2024 majority	36	–2.3568	0.1212	4.4015
EXIST2024 minority	38	–3.5089	0.0000	5.5672

Hard Evaluation. Table 9 presents the results for the hard-hard evaluation of Task 4. Out of the 50 systems submitted for this task, only 37 ranked above

Table 9. Results of Task 4 in the hard-hard evaluation (the best submission from each team).

Run	Rank	ICM-Hard	ICM-Hard Norm	$F1_{YES}$
EXIST2024 gold	0	0.9832	1.0000	1.0000
RoJiNG-CL_3 [20]	1	0.3182	0.6618	0.7642
I2C-Huelva_2	4	0.1313	0.5668	0.7241
DiTana-PV_2 [25]	6	0.1150	0.5585	0.7122
Victor-UNED_2	7	0.1028	0.5523	0.7154
MMICI_2	8	0.1014	0.5515	0.7261
I2C-Huelva_3	9	0.0987	0.5502	0.6933
NICA_1	11	0.0767	0.5390	0.7248
OppositionalOppotision_1	14	0.0494	0.5251	0.7168
Elias&Sergio_1	15	0.0433	0.5220	0.6979
ROCurve_3	19	0.0088	0.5045	0.6834
PINK_1	20	0.0076	0.5039	0.7044
RMIT-IR_2	22	–0.0123	0.4938	0.6726
Miqarn_1	26	–0.1159	0.4411	0.6632
CNLP-NITS-PP_1	27	–0.1234	0.4372	0.6699
Penta-ML_2	28	–0.1308	0.4335	0.6742
epistemologos_1	30	–0.1823	0.4073	0.5503
TokoAI_1	31	–0.1872	0.4048	0.5639
UMUTEAM_1	33	–0.2422	0.3768	0.6963
Umera Wajeed Pasha_1 [29]	36	–0.3083	0.3432	0.5956
TargaMarhuenda_1	37	–0.3535	0.3202	0.6487
EXIST2024 majority	39	–0.4038	0.2947	0.6821
DLRG_1	41	–0.4206	0.2861	0.6469
MIND_1 [21]	42	–0.4986	0.2465	0.5674
ALC-UPV-JD-2_1	43	–0.5446	0.2231	0.4878
dap-upv_1 [27]	44	–0.5737	0.2082	0.4188
AI Fusion_1	45	–0.6416	0.1737	0.4651
EXIST2024 minority	46	–0.6468	0.1711	0.0000
TheATeam_1	50	–0.6644	0.1621	0.4821
melialo-vcassan_2	51	–0.6644	0.1621	0.0281

the majority class baseline (all instances labeled as "NO"), while 47 systems surpassed the minority class baseline (all instances labeled as "YES"). Similar to the soft-soft evaluation, the results vary considerably, from 0.6618 ICM-Hard Norm for the best performing system (RoJiNG-CL_3) to 0.0876 (melialo-vcassan_1).

6.5 Task 5: Source Intention in Memes

Soft Evaluation. Table 10 presents the results for the classification of memes according to the intention of the author, with the outputs provided as the probabilities of the different classes. Only 15 runs were submitted for this task. While all the runs ranked above the minority class baseline (all instances labeled as "JUDGEMENTAL"), only 15 runs surpassed the majority class baseline (all instances labeled as "NO"). The results for this task are notably low, with the best team ("Victor-UNED_2") achieving only 0.3676 ICM-Soft Norm. This suggests that identifying whether a meme contains direct sexism or is judgmental is more difficult than identifying the intention behind a sexist tweet.

Table 10. Results of Task 5 in the soft-soft evaluation (the best submission from each team).

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
EXIST2024 gold	0	4.7018	1.0000	0.9325
Victor-UNED_2	1	–1.2453	0.3676	1.6235
MMICI_1	2	–1.2660	0.3654	1.4645
NICA_1	4	–1.5329	0.3370	1.4664
CNLP-NITS-PP_1	5	–1.5907	0.3308	1.5273
melialo-vcassan_2	6	–1.9847	0.2889	1.5211
I2C-Huelva_3	10	–2.7996	0.2023	3.9604
EXIST2024 majority	14	–5.0745	0.0000	5.5565
Penta-ML_3	15	–5.2668	0.0000	5.1547
EXIST2024 minority	18	–18.9382	0.0000	8.0245

Hard Evaluation. Table 11 presents the results for the hard-hard evaluation of Task 5. Out of the 19 systems submitted for this task, only 15 ranked above the majority class baseline (all instances labeled as "NO"), while 18 systems surpassed the minority class baseline (all instances labeled as "JUDGEMENTAL"). The results range from 0.4167 ICM-Hard Norm for the best performing system ("Victor-UNED_1") to 0.0000 for the worst performing systems, but are quite homogeneous among the top 5 systems.

Table 11. Results of Task 5 in the hard-hard evaluation (the best submission from each team).

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
EXIST2024 gold	0	1.4383	1.0000	1.0000
Victor-UNED_1	1	–0.2397	0.4167	0.3873
I2C-Huelva_2	2	–0.2535	0.4119	0.4761
NICA_1	6	–0.2881	0.3999	0.3837
MMICI_1	7	–0.3066	0.3934	0.4179
CNLP-NITS-PP_1	9	–0.3370	0.3829	0.4101
Penta-ML_3	11	–0.6123	0.2872	0.3841
TokoAI_1	14	–0.7263	0.2475	0.3716
melialo-vcassan_3	15	–0.7758	0.2303	0.3709
EXIST2024 majority	17	–1.0445	0.1369	0.1839
UMUTEAM_1	18	–1.1486	0.1007	0.2098
DLRG_1	20	–1.4891	0.0000	0.2530
EXIST2024 minority	21	–2.0637	0.0000	0.0697
epistemologos_1	22	–8.7012	0.0000	0.0557

6.6 Task 6: Sexism Categorization in Memes

Soft Evaluation. Table 12 presents the results for classifying memes based on the aspects of women being attacked, with outputs provided as class probabilities. Only 19 runs were submitted for this task. While all runs performed better than the minority class baseline (labeling all instances as "MISOGYNY-NON-SEXUAL-VIOLENCE"), only 11 runs exceeded the majority class baseline (labeling all instances as "NO"). The performance for this task was generally low, with the top team ("ROCurve_1") achieving an ICM-Soft Norm score of only 0.2462, which is significantly lower compared to the results for the same task when applied to tweets (Task 3).

Hard Evaluation. Finally, Table 13 presents the results for classifying memes based on the aspects of women being attacked, with outputs provided as a single class prediction. 22 runs were submitted for this task. Only 17 runs exceeded the majority class baseline (labeling all instances as "NO"), while 21 runs ranked above the minority class (all instances labeled as "MISOGYNY-NON-SEXUAL-VIOLENCE") The performance for this task was low, with the top team ("DiTana-PV_1") achieving an ICM-Soft Norm score of 0.3549.

Table 12. Results of Task 6 in the soft-soft evaluation (the best submission from each team).

Run	Rank	ICM-Soft	ICM-Soft Norm
EXIST2024 gold	0	9.4343	1.0000
ROCurve_1	1	–4.7893	0.2462
the gym nerds_2	2	–4.7942	0.2459
Elias&Sergio_1	5	–5.9160	0.1865
Victor-UNED_1	6	–6.4124	0.1602
CNLP-NITS-PP_1	8	–6.6782	0.1461
AI Fusion_1	10	–7.6282	0.0957
EXIST2024 majority	13	–9.8173	0.0000
dap-upv_1	14	–10.4213	0.0000
Penta-ML_2	15	–11.2593	0.0000
MMICI_1	19	–16.1248	0.0000
EXIST2024 minority	22	–50.0353	0.0000

Table 13. Results of Task 6 in the hard-hard evaluation (the best submission from each team).

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
EXIST2024 gold	0	2.4100	1.0000	1.0000
DiTana-PV_1	1	–0.6996	0.3549	0.4319
MMICI_1	3	–0.9863	0.2954	0.4342
ROCurve_1	4	–1.0089	0.2907	0.3639
Penta-ML_3	8	–1.3631	0.2172	0.3356
Elias&Sergio_1	11	–1.5276	0.1831	0.4321
Miqarn_1	13	–1.6216	0.1636	0.3211
CNLP-NITS-PP_1	14	–1.7920	0.1282	0.1587
ALC-UPV-JD-2_1	15	–1.8573	0.1147	0.2103
dap-upv_1	17	-1.9497	0.0955	0.2227
UMUTEAM_1	18	–1.9511	0.0952	0.3786
EXIST2024 majority	19	-2.0711	0.0703	0.0919
TargaMarhuenda_1	20	–2.0725	0.0700	0.1440
TheATeam_1	22	–2.3159	0.0195	0.1490
EXIST2024 minority	23	–3.3135	0.0000	0.0318
One-by-zero_1	25	–4.5910	0.0000	0.2304

7 Conclusions

The objective of the EXIST challenge is to encourage research on the automated detection and modeling of sexism in online environments, with a specific focus on social networks. The EXIST 2024 Lab held as part of CLEF attracted nearly 60 participant teams, and received a total of 412 runs. Participants adopted a wide range of approaches, including vision transformer models, data augmentation through automatic translation, data duplication, utilization of data from past EXIST editions, multilingual language models, Twitter-specific language models, and transfer learning techniques from domains like hate speech, toxicity, and sentiment analysis. While many systems opted for the traditional approach of providing only hard labels as outputs, a significant number of systems leveraged the multiple annotations available in the dataset, and provided soft outputs, proving that there is an increasing interest by the research community in developing systems able to deal with disagreements and with different perspectives.

For future editions of EXIST, we plan to expand our study in order to include additional communication channels and media formats, such as TikTok videos. By doing so, we aim to address the nuances and unique challenges presented by different formats, enhancing the robustness and applicability of research on automated sexism detection. Additionally, this expansion will allow us to capture a broader spectrum of online interactions and cultural contexts.

Acknowledgments. This work has been financed by the European Union (NextGenerationEU funds) through the "Plan de Recuperación, Transformación y Resiliencia", by the Ministry of Economic Affairs and Digital Transformation and by the UNED University. It has also been financed by the Spanish Ministry of Science and Innovation (project FairTransNLP (PID2021-124361OB-C31 and PID2021-124361OB-C32)) funded by MCIN/AEI/10.13039/501100011033 and by ERDF, EU A way of making Europe, and by the Australian Research Council (DE200100064 and CE200100005).

References

1. Amigó, E., Delgado, A.: Evaluating extreme hierarchical multi-label classification. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, vol. Volume 1: Long Papers, pp. 5809–5819. ACL, Dublin, Ireland (2022)
2. Aru, G., Emmolo, N., Piras, A., Marzeddu, S., Raffi, J., Passaro, L.C.: RoBEXedda: enhancing sexism detection in tweets for the EXIST 2024 challenge. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
3. Azadi, A., Ansari, B., Zamani, S.: Bilingual sexism classification: fine-tuned XLM-RoBERTa and GPT-3.5 few-shot learning. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
4. Barua, D.D., et al.: Penta ML at EXIST 2024: tagging sexism in online multimodal content with attention-enhanced modal context. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)

5. Billig, M.: Humour and hatred: the racist jokes of the Ku Klux Klan. Discourse Soc. **12**(3), 267–289 (2014)
6. Carrillo-Casado, Á., Román-Pásaro, J., Mata-Vázquez, J., Pachón-Álvarez, V.: I2C-UHU at EXIST 2024: transformer-based detection of sexism and source intention in memes using a learning with disagreement approach. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
7. Chakravarthi, B.R., et al.: Overview of shared task on multitask meme classification - unraveling misogynistic and trolls in online memes. In: Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, pp. 139–144 (2024)
8. Chulvi, B., Fontanella, L., Labadie-Tamayo, R., Rosso, P.: Social or individual disagreement? Perspectivism in the annotation of sexist jokes. In: Proceedings of the NLPerspectives 2023: 2nd Workshop on Perspectivist Approaches to Disagreement in NLP, co-locotaed with ECAI-2023 (2023)
9. Fan, S., Frick, R.A., Steinebach, M.: FraunhoferSIT@EXIST2024: leveraging stacking ensemble learning for sexism detection. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
10. Fang, Y.Z., Lee, L.H., Huang, J.D.: NYCU-NLP at EXIST 2024 – leveraging transformers with diverse annotations for sexism identification in social networks. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
11. Fersini, E., et al.: SemEval-2022 Task 5: multimedia automatic misogyny identification. In: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pp. 533–549 (2022)
12. Gasparini, F., Rizzi, G., Saibene, A., Fersini, E.: Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. Data Brief **44**, 108526 (2022)
13. Guerrero-García, M., Cerrejón-Naranjo, M., Mata-Vázquez, J., Pachón-Álvarez, V.: I2C-UHU at EXIST2024: learning from divergence and perspectivism for sexism identification and source intent classification. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
14. Hodson, G., Rush, J., MacInnis, C.C.: A joke is just a joke (except when it isn't): cavalier humor beliefs facilitate the expression of group dominance motives. J. Pers. Soc. Psychol. **99**(4), 660–682 (2010)
15. Jimenez-Martinez, M.P., Raygoza-Romero, J.M., Sánchez-Torres, C.E., Lopez-Nava, I.H., Montes-y Gómez, M.: Enhancing sexism detection in tweets with annotator-integrated ensemble methods and multimodal embeddings for memes. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
16. Keinan, R.: Sexism identification in social networks using TF-IDF embeddings, preproccessing, feature selection, word/Char N-grams and various machine learning models in Spanish and English. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
17. Khan, S., Pergola, G., Jhumka, A.: Multilingual sexism identification via fusion of large language models. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
18. Kirk, H.R., Yin, W., Vidgen, B., Röttger, P.: SemEval-2023 task 10: explainable detection of online sexism. In: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval) (2023)

19. Labadie-Tamayo, R., Chulvi, B., Rosso, P.: Everybody hurts, sometimes. Overview of HUrtful HUmour at IberLEF 2023: detection of humour spreading prejudice in Twitter. In: Procesamiento del Lenguaje Natural (SEPLN), pp. 383–395, No. 71 (2023)
20. Ma, J., Li, R.: RoJiNG-CL at EXIST 2024: sexism identification in memes by integrating prompting and fine-tuning. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
21. Maqbool, F., Fersini, E.: A contrastive learning based approach to detect sexism in memes. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
22. Maqbool, N.: Sexism identification in social networks: advances in automated detection – a report on the exist task at CLEF. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
23. Martinez, E., Cuadrado, J., Martinez-Santos, J.C., Puertas, E.: VerbaNex AI at CLEF EXIST 2024: detection of online sexism using transformer models and profiling techniques. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
24. Mendiburo-Seguel, A., Ford, T.E.: The effect of disparagement humor on the acceptability of prejudice. Current Psychology: A Journal for Diverse Perspectives on Diverse Psychological Issues, pp. No Pagination Specified–No Pagination Specified (2019)
25. Menárguez Box, A., Torres Bertomeu, D.: DiTana-PV at sEXism identification in social networks (EXIST) tasks 4 and 6: the effect of translation in sexism identification. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
26. Naebzadeh, A., Nobakhtian, M., Eetemadi, S.: NICA at EXIST CLEF tasks 2024. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
27. Obrador Reina, M., García Cucó, A.: LightGMB for sexism identification in memes. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
28. Pan, R., García Díaz, J.A., Bernal Beltrán, T., Valencia-Garcia, R.: UMUTeam at EXIST 2024: multi-modal identification and categorization of sexism by feature integration. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
29. Pasha, U.W.: Multilingual sexism detection in memes, a CLIP-enhanced machine learning approach. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
30. Petrescu, A., Truică, C.O., Apostol, E.S.: Language-based mixture of transformers for EXIST2024. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
31. Plaza, L., et al.: Overview of EXIST 2023 – learning with disagreement for sexism identification and characterization (Extended Overview). In: Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum (2023)
32. Plaza, L., et al.: Overview of EXIST 2024 – learning with disagreement for sexism identification and characterization in social networks and memes (Extended Overview). In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
33. Plaza, L., et al.: Overview of EXIST 2023 – learning with disagreement for sexism identification and characterization (Extended Overview). In: Aliannejadi, M.,

Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), vol. 497, pp. 813–854. CEUR Working Notes (2023)
34. Quan, L.M., Thin, D.V.: Sexism identification in social networks with generation-based approach. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
35. Rizzi, G., Gimeno-Gómez, D., Fersini, E., Martínez-Hinarejos, C.D.: PINK at EXIST2024: a cross-lingual and multi-modal transformer approach for sexism detection in memes. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
36. Rodríguez-Sánchez, F., et al.: Overview of EXIST 2021: sexism identification in social networks. Procesamiento del Lenguaje Natural **67**, 195–207 (2021)
37. Rodríguez-Sánchez, F., et al.: Overview of EXIST 2022: sexism identification in social networks. Procesamiento del Lenguaje Natural **69**, 229–240 (2022)
38. Ruiz, V., Carrillo-de-Albornoz, J., Plaza, L.: Concatenated transformer models based on levels of agreements for sexism detection. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
39. Shah, A., Gokhale, A.: Team Aditya at EXIST 2024 – detecting sexism in multilingual tweets using contrastive learning approach. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
40. Shanbhag, A., Jadhav, S., Date, A., Joshi, S., Sonawane, S.: The wisdom of weighing: stacking ensembles for a more balanced sexism detector. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
41. Shifat, F.T., et al.: Penta-NLP at EXIST 2024 Task 1–3: sexism identification, source intention, sexism categorization in tweets. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
42. Shimi, G., Mahibha, J., Thenmozhi, D.: Automatic classification of gender stereotypes in social media post. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
43. Siino, M., Tinnirello, I.: Prompt engineering for identifying sexism using GPT mistral 7B. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
44. Smith, T., Nie, R., Trippas, J., Spina, D.: RMIT-IR at EXIST lab at CLEF 2024. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
45. Murari Sreekumar, S.K., Thenmozhi, D., Gopalakrishnan, S., Swaminathan, K.: Sexism identification in tweets using traditional machine learning approaches. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
46. Tavarez-Rodríguez, J., Sánchez-Vega, F., Rosales-Pérez, A., López-Monroy, A.P.: Better together: LLM and neural classification transformers to detect sexism. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
47. Uma, A., et al.: SemEval-2021 task 12: learning with disagreements. In: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pp. 338–347. Association for Computational Linguistics, Online, August 2021
48. Usmani, M., Siddiqui, R., Rizwan, S., Khan, F., Alvi, F., Samad, A.: Sexism identification in tweets using BERT and XLM – Roberta. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)

49. Vetagiri, A., Mogha, P., Pakray, P.: Cracking down on digital misogyny with MULTILATE a MULTImodaL hATE detection system. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
50. Villarreal-Haro, K., Sánchez-Vega, F., Rosales-Pérez, A., López-Monroy, A.P.: Stacked reflective reasoning in large neural language models. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)

Intelligent Disease Progression Prediction: Overview of iDPP@CLEF 2024

Giovanni Birolo[1], Pietro Bosoni[2], Guglielmo Faggioli[3(✉)], Helena Aidos[4], Roberto Bergamaschi[2], Paola Cavalla[1,5], Adriano Chiò[1], Arianna Dagliati[2], Mamede de Carvalho[4], Giorgio Maria Di Nunzio[3], Piero Fariselli[1], Jose Manuel García Dominguez[6], Marta Gromicho[4], Alessandro Guazzo[3], Enrico Longato[3], Sara C. Madeira[4], Umberto Manera[1], Stefano Marchesin[3], Laura Menotti[3], Gianmaria Silvello[3], Eleonora Tavazzi[7], Erica Tavazzi[3], Isotta Trescato[3], Martina Vettoretti[3], Barbara Di Camillo[3], and Nicola Ferro[3]

[1] University of Turin, Turin, Italy
{giovanni.birolo,paola.cavalla,adriano.chio,piero.fariselli, umberto.manera}@unito.it
[2] University of Pavia, Pavia, Italy
{pietro.bosoni,arianna.dagliati}@unipv.it, roberto.bergamaschi@mondino.it
[3] University of Padua, Padua, Italy
{guglielmo.faggioli,alessandro.guazzo,enrico.longato,stefano.marchesin, laura.menotti,gianmaria.silvello,erica.tavazzi,isotta.trescato, barbara.dicamillo,nicola.ferro}@phd.unipd.it, martina.vettoretti@unipd.it
[4] University of Lisbon, Lisbon, Portugal
{haidos,sacmadeira}@fc.ul.pt, mamedemg@mail.telepac.pt, mgromichosilva@medicina.ulisboa.pt
[5] "Città della Salute e della Scienza", Turin, Italy
[6] Gregorio Marañon Hospital in Madrid, Madrid, Spain
jgarciadominguez@salud.madrid.org
[7] IRCCS Foundation C. Mondino in Pavia, Pavia, Italy

Abstract. Multiple Sclerosis (MS) and Amyotrophic Lateral Sclerosis (ALS) are two neurodegenerative diseases that cause progressive or alternating neurological impairments in motor, sensory, visual, and cognitive functions. Patients affected by these diseases undergo the physical, psychological, and economic burdens of hospital stays and home care while facing uncertainty. A possible aid to patients and clinicians might come from AI tools that can preemptively identify the need for intervention and suggest personalized therapies during the progression of these diseases.

The objective of iDPP@CLEF is to develop automatic approaches based on AI that can be used to describe the progression of these two neurodegenerative diseases, with the final goal of allowing patient stratification as well as the prediction of the disease progression, to help clinicians in assisting patients in the most timely manner.

G. Birolo, P. Bosoni and G. Faggioli—These authors contributed equally.

L. Goeuriot et al. (Eds.): CLEF 2024, LNCS 14959, pp. 118–139, 2024.
https://doi.org/10.1007/978-3-031-71908-0_6

iDPP@CLEF 2024 follows the two prior editions, iDPP@CLEF 2022 and 2023. iDPP@CLEF 2022 focused on ALS progression prediction and approaches of explainable AI for the task. iDPP@CLEF 2023 built upon iDPP@CLEF 2022 by extending the datasets provided during the previous edition with environmental data. Additionally, the 2023 edition of iDPP@CLEF introduced a new task focused on the progression prediction of MS. In this edition, we extended the MS dataset of iDPP@CLEF 2023 with environmental data. Furthermore, we introduced two new ALS tasks, focused on predicting the progression of the disease using data obtained from wearable devices, making it the first iDPP edition that uses prospective data collected directly from the patients involved in the BRAINTEASER project.

1 Introduction

Amyotrophic Lateral Sclerosis (ALS) and *Multiple Sclerosis (MS)* are two severe and extremely impactful diseases that cause progressive neurological impairment in people living with them. Typically, for both diseases, the progression is heterogeneous, determining a large variability in several aspects including the treatment of the patients, the outcome, the quality of life, and, in general, their needs. This variability represents a challenge not only for patients but also for clinicians and caregivers alike. Indeed, patients with ALS tend to require, at a certain point of the progression of their disease, some specific treatment, such as *NonInvasive Ventilation (NIV)* or *Percutaneous Endoscopic Gastrostomy (PEG)*. Similarly, patients living with MS tend to undergo impairing relapses that may cause severe drops in their quality of life. Therefore, it would beneficial to know beforehand the needs of a person affected by one of these diseases. Nevertheless, due to their heterogeneity, it is challenging to develop effective prognostic tools. This motivates the importance of developing automatic tools to aid clinicians in their decision-making in all phases of disease progression and facilitate personalized therapeutic choices. In particular, when developing new automatic predictive approaches based on *Artificial Intelligence (AI)*, researchers need a proper framework that allows for designing and evaluating approaches for different tasks, such as:

- stratifying patients according to their phenotype all over the disease evolution;
- predicting the progression of the disease in a probabilistic, time-dependent way;
- describing better and in an explainable fashion the mechanisms underlying MS and ALS diseases.

Nonetheless, it is of uttermost importance that such approaches are based on shared resources that allow for proper benchmarking and comparable and reproducible experimentation. The *Intelligent Disease Progression Prediction at CLEF (iDPP@CLEF)* lab[1] aims to provide an evaluation infrastructure for

[1] https://brainteaser.health/open-evaluation-challenges/.

the development of such AI algorithms. Differently from previous efforts in this domain, iDPP@CLEF systematically addresses issues related to the application of AI in clinical practice for ALS and MS. Apart from defining risk scores based on the probability of events occurring in the short or long term, iDPP@CLEF also deals with providing clinicians with structured and understandable data. iDPP@CLEF 2024 [2] is the last iteration of an evaluation cycle of three challenges aimed at fostering reproducible and comparable evaluation of AI based approaches to predict the progression of ALS and MS. The first edition, iDPP@CLEF 2022 focused on ALS, asking participants to predict the probability that patients would incur the need for specific medical treatments based on their medical history. The second edition iDPP@CLEF 2023 focused on extending the dataset of iDPP@CLEF 2022 with environmental data, to determine the impact that the environment might have on the needs of the patients. Furthermore, a new task based on predicting the risk of worsening of MS patients was proposed. This final edition extends iDPP@CLEF 2023 by providing environmental data for patients affected by MS, to measure the impact that pollution and the external environment can have on the progression of MS. Furthermore, two new tasks have been proposed in iDPP@CLEF 2024. These tasks required participants to predict the progression of ALS, measured according to the ALSFRS-R scale, based on the clinical history of the patients, as well as measurements obtained via wearable devices and sensors.

The paper is organized as follows: Sect. 2 presents related challenges; Sect. 3 describes its tasks; Sect. 4 discusses the developed dataset; Sect. 5 explains the setup of the Lab and introduces the participants; Sect. 6 introduces the evaluation measures adopted to score the runs; Sect. 7 analyzes the experimental results for the different tasks; finally, Sect. 8 draws some conclusions and outlooks some future work.

2 Related Challenges

There have been no other labs on this or similar topics within CLEF before the start of iDPP@CLEF. iDPP@CLEF 2022 and 2023 were the first two iterations of the Lab and the current is the third.

While no major challenges – besides iDPP@CLEF 2023 – regarding MS have been carried out yet, more interest has been shown toward ALS. In particular, three major challenges were organized on this topic: the DREAM 7 ALS Prediction challenge[2] in 2012 and the DREAM ALS Stratification challenge[3] in 2015 and a Kaggle challenge[4] in 2021. The DREAM 7 ALS Prediction challenge consisted of using 3 months of ALS clinical trial information (months 0–3) to predict the future progression of the disease (months 3–12), expressed as the slope of change in *ALS Functional Rating Scale Revisited (ALSFRS-R)* [4]. Later on, the DREAM ALS Stratification challenge [9] required participants to stratify

[2] https://dreamchallenges.org/dream-7-phil-bowen-als-prediction-prize4life/.
[3] https://dx.doi.org/10.7303/syn2873386.
[4] https://www.kaggle.com/alsgroup/end-als.

ALS into subgroups based on their characteristics, to understand patient profiles better and provide personalized ALS treatments. Finally, the Kaggle challenge employed clinical and genomic data to obtain a better understanding of the mechanisms underlying ALS and determine why some people with ALS tend to have a faster progression of the disease compared to others.

At the current time, most of the datasets used to evaluate AI algorithms for MS are based on closed and proprietary datasets. In this sense iDPP@CLEF paved the way for a reproducible and effectively open science in the research domain of the AI used for predicting the progression of MS.

2.1 iDPP@CLEF 2022

iDPP@CLEF 2022[5] [7,8] was the first edition of the Lab and concerned exclusively the ALS disease progression prediction. Being the pilot Lab, a large share of effort was devoted to understanding the challenges and limitations linked to the shared evaluation campaigns, when it comes to AI applied in the medical domain. iDPP@CLEF 2022 was organized into 3 tasks:

- **Pilot Task 1 - Ranking Risk of Impairment**: The focus of the first task of iDPP@CLEF 2022 was on ranking patients based on the risk of impairment, defined as the need for specific medical treatments, such as NIV, PEG, or death. Participants were given information on the motor functioning of the patients in time, measured according to the ALSFRS-R scale [4], and were asked to rank patients based on the time-to-event risk of experiencing impairment in each specific domain.
- **Pilot Task 2 - Predicting Time of Impairment**: It refined Task 1 by asking participants to predict when specific impairments will occur (i.e. in the correct time window). The task focused on assessing model calibration in terms of the ability of the proposed algorithms to estimate the probability of an event close to the true probability within a specified time window.
- **Position Paper Task 3 - Explainability of AI algorithms**: The task focused on the evaluation and discussion of AI-based explainable frameworks for intelligent disease progression prediction able to explain the multivariate nature of the data and the model predictions.

One of the major outputs of iDPP@CLEF 2022 were the three datasets released. In particular, the datasets contain data for the prediction of specific events related to ALS. Such datasets are fully anonymized retrospective details about 2250 real patients. The patients were recruited from two medical institutions in Turin, Italy, and Lisbon, Portugal. The datasets contain static data about patients (e.g., age, onset date, gender) and event data (i.e. 18,512 ALSFRS-R questionnaires and 4,015 spyrometries). six groups participated in iDPP@CLEF 2022 and submitted a total of 120 runs.

[5] https://brainteaser.health/open-evaluation-challenges/idpp-2022/.

2.2 iDPP@CLEF 2023

Similarly to iDPP@CLEF 2022, also iDPP@CLEF 2023[6] [5,6] were organized into three tasks, focusing on either ALS or MS. More in detail, Task 1 and Task 2 of iDPP@CLEF 2023 concerned MS, while Task 3 built upon iDPP@CLEF 2022 and extended the ALS tasks of the previous iteration of the Lab. To summarize iDPP@CLEF 2023 tasks:

- **Task 1: Predicting Risk of Disease Worsening (MS)** This task focused on predicting the probability that, given the history of the patient, they would undergo a worsening, according to two different definitions of worsening.
- **Task 2: Predicting Cumulative Probability of Worsening (MS)** The second task had a similar objective to Task 1, with the major difference that, instead of predicting the risk at an absolute level, participants were required to predict the cumulative probability of worsening over 10 years.
- **Task 3: Position Papers on the Impact of Exposition to Pollutants (ALS)** The third task extended the first task of iDPP@CLEF 2022 and concerned the ranking of the patients based on the risk of impairment. The major difference to iDPP@CLEF 2022 was that participants were given environmental data to determine if such data was a good predictor of the risk of impairment.

iDPP@CLEF 2023 extended the iDPP@CLEF 2022 datasets with two datasets for MS. In particular, such datasets contained static data about patients, MS-related details (e.g., the EDSS score, results of MRIs, evoked potentials measures), and a label indicating if the patient underwent a worsening, based on the worsening definitions of Task 1 and Task 2. Ten teams submitted a total of 163 runs at the end of iDPP@CLEF 2023.

3 Tasks

In the remainder of this section, we describe each task in more detail.

3.1 Task 1: Predicting ALSFRS-R Score from Sensor Data (ALS)

Task 1 focuses on predicting the twelve scores of the ALSFRS-R (ALS Functional Rating Scale - Revised), assigned by medical doctors roughly every three months, from the sensor data collected via the app. The ALSFRS-R is a somewhat "subjective" evaluation usually performed by a medical doctor. This task aims to answer an open question in the research community, i.e., whether the ALSFRS-R scores can be derived from objective factors.

Participants were given the ALSFRS-R questionnaire at the first visit, including the scores for each question and the time (number of days from diagnosis) when the questionnaire was taken; moreover, they were also provided with the

[6] https://brainteaser.dei.unipd.it/challenges/idpp2023/.

time of the second visit (number of days from diagnosis) and all the sensor data up to the time of the second visit.

Participants had to predict the values of the ALSFRS-R sub-scores at the second visit.

3.2 Task 2: Predicting Patient Self-assessment Score from Sensor Data (ALS)

The second task concerning ALS focuses on predicting the self-assessment scores assigned by patients from the sensor data collected via the app. Self-assessment scores correspond to each of the ALSFRS-R scores, but while the latter are assigned by medical doctors during visits, the self-assessment scores are assigned by patients themselves using the app.

If the self-assessment performed by patients, which occurs more frequently than the assessments performed by medical doctors every three months or so, can be reliably predicted by sensor and app data, we can imagine a proactive application that monitors the sensor data and alerts the patient if an assessment is needed.

Participants were given the first set of self-assessed scores along with the time (number of days from diagnosis) at which the questionnaire was taken; furthermore, they were also provided with the time of the second auto-evaluation (number of days from diagnosis) and all the sensor data up to the time of the second auto-evaluation. Participants had to predict the values of the self-assessed scores at the second auto-evaluation, which occurs one or two months after the first one.

3.3 Task 3: Predicting Relapses from EDSS Sub-scores and Environmental Data (MS)

The third task focuses on predicting a relapse using environmental data and EDSS (Expanded Disability Status Scale) sub-scores. This task allows us to assess if exposure to different pollutants is a useful variable in predicting a relapse.

Participants were asked to predict the week of the first relapse after the baseline considering environmental data based on a weekly granularity, given the status of the patient at the baseline, which is the first visit available in the considered time span (after January 1, 2013). For each patient, the date of the baseline will be week 0 and all the other weeks will be relative to it.

Participants were given all the environmental data about a patient, i.e. also observations which may happen after the relapse to be predicted. All the patients are guaranteed to experience, at least, one relapse after the baseline.

4 Dataset

For iDPP@CLEF 2024 we release three datasets: two completely new datasets for ALS and an extension of the iDPP@CLEF 2023 dataset concerning MS. More in detail, the two new ALS datasets comprise a common training part with 52 training patients, whose ALSFRS-R scores were both annotated by the clinicians and self-assessed. Concerning the test sets, 21 and 11 patients were included in them for Task 1 and Task 2 respectively. Regarding MS, the part of the dataset concerning static variables and MS-related information is the same as the one used for iDPP@CLEF 2023. The major improvement regards environmental data that have been added to the dataset.

4.1 Tasks 1 and 2: ASL Dataset with Clinical or Self-assessed ALSFRS-R

The datasets for Task 1 and Task 2 were collected from ALS-diagnosed patients recruited during the BRAINTEASER project from three centers in Lisbon, Madrid, and Turin. At recruitment, patients were given a commercial fitness tracker (the Garmin VivoActive 4 smartwatch), and data from its sensors was collected during a follow-up period with a median duration of 270 days. Patients were encouraged to wear the watch as much as they were comfortable with, ideally all the time, both while awake and sleeping. Each day of data for each patient was summarized into a vector of 90 statistics related to heart rate and beat-to-beat interval, respiration rate, and nocturnal pulse oximetry. Sensor data was not available every day for each patient.

During the same period, disease progression was assessed by their clinician using the ALSFRS-R questionnaire (roughly every three months, following standard clinical practice). Patients also used the same questionnaire to self-assess their progression through a smartphone app developed specifically by the BRAINTEASER project. They were prompted for the assessment once per month, though the actual frequency varied and depended on patient compliance.

Creation of the Datasets. Patients with insufficient data were excluded from the challenge dataset. Specifically, this included those with less than three months of follow-up data, those with more than 50% of sensor data missing, and those without at least two clinical or self-assessed ALSFRS-R evaluations. After applying these criteria, a dataset of 83 patients was obtained, with a median of 254 days of sensor data per patient. These patients and their data were then divided into a training group (common to both Tasks 1 and 2) and two task-specific testing groups.

Split Into Training and Test. The patients were split into three groups:

training patients with at least two clinical and two self-assessed ALSFRS-R evaluations;

test-ct patients with at least two clinical but without two self-assessed ALSFRS-R evaluations;

test-app patients with at least two self-assessed but without two clinical ALSFRS-R evaluations.

The training set thus included 52 patients with a median of 3.5 clinical and 5 self-assessed ALSFRS-R evaluations (189 and 301 in total, respectively). The test-ct set (the test set for Task 1) included 21 patients, whose first clinical ALSFRS-R evaluations were included as features and the second evaluations were the prediction target. The test-app set (the test set for Task 2) included 11 patients and was built in the same way using the self-assessed ALSFRS-R evaluations. The full available sensor data for all patients was included in both the training and test datasets, while only the clinical (resp. self-assessed) ALSFRS-R evaluations were included for Task 1 (resp. Task 2).

4.2 Task 3: MS Dataset

The dataset used for Task 3 in iDPP@CLEF 2024 is structured similarly to those from iDPP@CLEF 2023, though some features (e.g., evoked potentials, MRIs) were not included, and certain records have been filtered based on the purpose of the task.

Updates over IDPP@CLEF 2023. In the 2024 dataset, EDSS data after January 1, 2013 (aligned with the start of environmental data collection), were filtered, and patients without EDSS follow-ups were removed. Additionally, patients who did not experience a relapse after their first non-filtered EDSS follow-up (i.e., the baseline for each patient) were excluded.

The dataset has been expanded to incorporate environmental data, which includes information on patients' exposure to various air pollutants identified as significant public health risks in the latest World Health Organization (WHO) global air quality guidelines [15], such as particulate matter (PM) - encompassing both $PM_{2.5}$ (particles with an aerodynamic diameter of 2.5 µm or less) and PM_{10} (particles with an aerodynamic diameter of 10 µm or less) - as well as ozone (O_3), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and several weather factors (including wind speed, relative humidity, sea level pressure, global radiation, precipitation, and average, minimum, and maximum temperatures).

Air pollutant data from public monitoring stations were collected daily from the European Air Quality Portal using the DiscoMap tool[7]. The geographical coordinates (longitude and latitude) of each monitoring station were matched to specific postcodes, identifying the nearest station to each patient's residence postcode. Instead, weather data were gathered daily from the European Climate Assessment and Dataset station network, which provides access to the E-OBS dataset, a daily gridded land-only observational dataset over Europe[8]. Each grid was matched with the nearest monitoring station using Euclidean distance based on geographical coordinates. This approach ensured that air pollution and weather data were aligned with the same spatial and temporal granularity. Daily environmental measurements were aggregated into weekly averages from

[7] https://discomap.eea.europa.eu/Index/.
[8] https://www.ecad.eu/download/ensembles/download.php.

each patient's baseline. As additional features, the number of days per week spent over the respective WHO recommended air quality guideline levels for short-term (24 h) exposure was computed for each air pollutant.

Finally, a subset of 380 MS patients from Turin and Pavia research centers was selected for Task 3 in iDPP@CLEF 2024, compared to 550 patients for Task 1 and 638 for Task 2 in iDPP@CLEF 2023. The resulting MS dataset[9] includes static variables, with demographic and clinical information, EDSS scores with corresponding Functional System (FS) sub-scores, environmental measurements, and the outcome time, representing the week of the first relapse occurrence after the baseline for each patient. EDSS visits occur between the baseline and the first relapse, while environmental measurements span from January 1, 2013, to 2023. It is important to note that environmental data may have gaps due to data availability.

Split Into Training and Test. The dataset was split into a training set (70%) and a test set (30%), with subjects stratified by outcome time to ensure an even distribution across both sets. The distribution of static data, including demographic and clinical information, and EDSS were verified to be similar in both training and test sets. Additionally, since environmental exposure is considered, the distribution of patients from the two clinical centres and their residence classification (Cities, Rural Areas, and Towns) was checked to be balanced.

Statistical tests, including the Kruskal-Wallis test for continuous variables and the Chi-squared test for categorical and ordinal variables, were performed to assess the appropriateness of the stratification. Special attention was given to sparsely observed levels in categorical variables to ensure rare levels appeared only in the training set if at all. Table 1 provides a comparison of variable distributions between the training and test sets, confirming that the split meets the best-practice quality standards.

5 Lab Setup and Participation

In the remainder of this section, we detail the guidelines the participants had to comply with to submit their runs and the submissions received by iDPP@CLEF.

5.1 Guidelines

Participating teams should satisfy the following guidelines:

- The runs should be submitted in the textual format described below;
- Each group can submit a maximum of 30 runs for each of Task 1 and Task 2 and Task 3.

[9] https://brainteaser.dei.unipd.it/challenges/idpp2024/assets/other/ms/ms-variables-description.txt.

Table 1. Comparison between training and test populations for MS task. Continuous variables are presented as median (interquartile range); categorical variables as count (percentage on available data), for each level.

Variable	Level	Levels Training	Levels Test
Sex	Female	148 (74.37%)	54 (66.67%)
	Male	51 (25.63%)	27 (33.33%)
Ethnicity	Caucasian	181 (90.96%)	77 (95.06%)
	Hispanic	2 (1.00%)	–
	Black African	2 (1.00%)	–
	NA	14 (7.04%)	4 (4.94%)
Residence classification	Cities	53 (26.63%)	20 (24.69%)
	Rural Area	52 (26.13%)	22 (27.16%)
	Towns	94 (47.24%)	39 (48.15%)
Centre	Pavia	129 (64.82%)	58 (71.61%)
	Turin	70 (35.18%)	23 (28.39%)
Occurrence of MS in pediatric age	FALSE	176 (88.44%)	77 (95.06%)
	TRUE	23 (11.56%)	4 (4.94%)
Age at onset	median (IQR)	28 (22-36)	30 (24-34)
Age at baseline	median (IQR)	38 (31-47)	38 (33-47)
Diagnostic delay	median (IQR)	12 (4-47)	12 (3-28)
Spinal cord symptom	FALSE	143 (71.86%)	54 (66.67%)
	TRUE	56 (28.14%)	27 (33.33%)
Brainstem symptom	FALSE	146 (73.37%)	57 (70.37%)
	TRUE	53 (26.63%)	24 (29.63%)
Eye symptom	FALSE	148 (74.37%)	59 (72.84%)
	TRUE	51 (25.63%)	22 (27.16%)
Supratentorial symptom	FALSE	140 (70.35%)	50 (61.73%)
	TRUE	59 (29.65%)	31 (38.27%)
Other symptoms	FALSE	197 (99.00%)	80 (98.77%)
	Sensory	1 (0.50%)	1 (1.23%)
	Epilepsy	1 (0.50%)	–
EDSS	median (IQR)	2.0 (1.5-3.0)	2.0 (1.5-3.5)
	NA	3 (0.36%)	0 (0.00%)
Outcome time	median (IQR)	59 (24-122)	53 (25-130)

Task 1 Run Format. Runs should be submitted as a text file (.txt) with the following format:

```
10061925618906738677 1 2 3 4 1 2 3 4 1 2 3 4 upd_T1_myDesc
10160033396142711519 1 2 3 4 1 2 3 4 1 2 3 4 upd_T1_myDesc
10287479530859953248 1 2 3 4 1 2 3 4 1 2 3 4 upd_T1_myDesc
12398828804459792214 1 2 3 4 1 2 3 4 1 2 3 4 upd_T1_myDesc
10038199677222038201 1 2 3 4 1 2 3 4 1 2 3 4 upd_T1_myDesc
...
```

where:

- Columns are separated by a white space;
- The first column is the patient ID, an hashed version of the original patient ID (should be considered just as a string);
- Columns from 2 to 13 represent the predicted ALSFRS-R sub-score. Each column corresponds to an ALSFRS-R question, e.g. column 2 to Q1, column 3 to Q2, and so on). Each values is expected to be integer in the range [0, 4];
- The last column is the run identifier, according to the format described below. It must uniquely identify the participating team and the submitted run.

It is important to include all the columns and have a white space delimiter between the columns. No specific ordering is expected among patients (rows) in the submission file.

Task 2 Run Format. Runs should be submitted as a text file (.txt) with the following format:

```
10061925618906738677 1 2 3 4 1 2 3 4 1 2 3 4 upd_T1_myDesc
10160033396142711519 1 2 3 4 1 2 3 4 1 2 3 4 upd_T1_myDesc
10287479530859953248 1 2 3 4 1 2 3 4 1 2 3 4 upd_T1_myDesc
12398828804459792214 1 2 3 4 1 2 3 4 1 2 3 4 upd_T1_myDesc
10038199677222038201 1 2 3 4 1 2 3 4 1 2 3 4 upd_T1_myDesc
...
```

where:

- Columns are separated by a white space;
- The first column is the patient ID, an hashed version of the original patient ID (should be considered just as a string);
- Columns from 2 to 13 represent the predicted self-assessd sub-score. Each column corresponds to an ALSFRS-R question, e.g. column 2 to Q1, column 3 to Q2, and so on). Each values is expected to be integer in the range [0, 4];
- The last column is the run identifier, according to the format described below. It must uniquely identify the participating team and the submitted run.

It is important to include all the columns and have a white space delimiter between the columns. No specific ordering is expected among patients (rows) in the submission file.

Task 3 Run Format. Runs should be submitted as a text file (.txt) with the following format:

```
10061925618906738677 10 upd_T3_myDesc
10160033396142711519 47 upd_T3_myDesc
10287479530859953248 13 upd_T3_myDesc
12398828804459792214 1 upd_T3_myDesc
10038199677222038201 9 upd_T3_myDesc
...
```

where:

- Columns are separated by a white space;
- The first column is the patient ID, a hashed version of the original patient ID (should be considered just as a string);
- The second column is the predicted week at which the first relapse after the baseline happens. The value is expected to be an integer starting from 1;
- The third column is the run identifier, according to the format described below. It must uniquely identify the participating team and the submitted run.

It is important to include all the columns and have a white space delimiter between the columns. No specific ordering is expected among patients (rows) in the submission file.

Submission Upload. Runs should be uploaded to the repository provided by the organizers. Following the repository structure discussed above, for example, a run submitted for the first task should be included in `submission/task1`.

Runs should be uploaded using the following name convention for their identifiers: `<teamname>_T<1|2|3>_<freefield>`, where:

- `teamname` is the name of the participating team;
- `T<1|2|3>` is the identifier of the task the run is submitted to, e.g. T1 for Task 1;
- `freefield` is a free field that participants can use as they prefer to further distinguish among their runs. Please, keep it short and informative.

For example, a complete run identifier may look like `upd_T1_myDesc`, where:

- `upd` is the University of Padua team;
- `T1` means that the run is submitted for Task 1;
- `myDesc` suggests an appropriate description for the run.

The name of the text file containing the run must be the identifier of the run followed by the txt extension. In the above example `upd_T1_myDesc.txt`

5.2 Participants

A total of 28 teams registered to iDPP@CLEF 2024, out of which 8 teams were able to submit one run in at least one task. Table 2 reports the details about teams that managed to submit at least one run. Furthermore, Table 3 outlines in which tasks each team participated in and how many runs they were able to submit. In total, 97 runs were submitted to iDPP@CLEF 2024. The most participated task was Task 1 with 59 runs and 6 teams participating. Subsequently, Task 2 had 31 runs submitted by 6 different teams. Finally, only two teams participated in task 3, with a total of 7 runs submitted. The most prolific participant was UNIPD, with a total of 20 runs.

Table 2. Teams participating in iDPP@CLEF 2024.

Team Name	Affiliation	Country	Repository	Paper
BIT.UA	IEETA/DETI, LASI, University of Aveiro	Portugal	https://bitbucket.org/brainteaser-health/idpp2024-bitua	Silva and Oliveira [14]
CompBiomedUniTO	University of Torino	Italy	https://bitbucket.org/brainteaser-health/idpp2024-compbiomedunito	Barducci et al. [1]
FCOOL	LASIGE, Faculty of Sciences, University of Lisbon	Portugal	https://bitbucket.org/brainteaser-health/idpp2024-fcool	Martins et al. [11]
iDPPExplorers	Georgia Institute of Technology, Atlanta, GA	United States	https://bitbucket.org/brainteaser-health/idpp2024-idppexplorers	Metha et al. [12]
Mandatory	University of Bucharest	Romania	https://bitbucket.org/brainteaser-health/idpp2024-mandatory	—
Stefagroup	University of Pavia, BMI lab"Mario Stefanelli"	Italy	https://bitbucket.org/brainteaser-health/idpp2024-stefagroup	Bosoni et al. [3]
UBCS	University of Botswana	Botswana	https://bitbucket.org/brainteaser-health/idpp2024-ubcs	Okere et al. [13]
UNIPD	University of Padova	Italy	https://bitbucket.org/brainteaser-health/idpp2024-unipd	Martinello et al. [10]

6 Evaluation Measures

In both Tasks 1 and 2, the prediction targets were the future scores of the ALSFRS-R evaluation, which are integers in the [0-4] range. Since the scores are discrete, we could have framed the predictive task as a classification problem. However, we opted for a regression problem to be able to penalize larger errors more (e.g., with a target score of 3, predicting 1 should be worse than predicting 2). Task 3, where the target was the week of the relapse, was also framed quite naturally as a regression task for similar reasons. Thus, we evaluated all tasks using the same two state-of-the-art evaluation measures to assess the performance of regression models: the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE). The formulas for RMSE and MAE are shown in Eq. 1 and Eq. 2, respectively, where n represents the number of observations, y_i is the actual value of the dependent variable for the i-th observation, and $\hat{y}_i$ is the predicted value of the dependent variable for the i-th observation.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

Table 3. Number of runs submitted by each participant team in iDPP@CLEF 2024

	Task 1 (ALS)	Task 2 (ALS)	Task 3 (MS)	Total
BIT.UA	7	7	—	14
CompBiomedUniTO	1	1	—	2
FCOOL	9	9	—	18
iDPPExplorers	15	—	—	15
Mandatory	19	—	—	19
Stefagroup	—	—	3	3
UBCS	—	6	—	6
UNIPD	8	8	4	20
Total	59	31	7	97

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{2}$$

Both metrics can explain the performance of a model in an interpretable manner since their units are the same as the target variable (e.g., weeks); together, they can provide a comprehensive evaluation of the three prediction tasks, with smaller values indicating better simulation results.

The RMSE measures how much, on average, the model's predictions deviate from the actual values. By squaring the errors before averaging them, RMSE gives higher weight to large errors. MAE represents the average absolute difference between actual and predicted values. Unlike RMSE, MAE treats all errors equally, regardless of their magnitude. Therefore, it provides a clear representation of the average error, is less sensitive to outliers, but does not emphasize large errors as much as RMSE.

7 Results

For each task, we report the analysis of the performance of the runs submitted by the Lab's participants according to the measures described in Sect. 6.

7.1 Task 1: Predicting ALSFRS-R Score from Sensor Data (ALS)

Clinicians monitor ALS progression through frequent visits, typically every two to three months, to promptly detect any worsening of symptoms. Consequently, ALSFRS-R scores usually remain fairly stable between these appointments, making the most recent score a reliable predictor for the next assessment. While some deterioration in at least one score is not uncommon, using the last observed value as a predictive measure is both simple and effective, as most scores will not change. This approach is particularly useful for bulbar and respiratory scores,

Table 4. For both MAE and RMSE, we report the average error across all twelve ALSFRS-R scores, the average standard deviation (computed by bootstrapping the test set one thousand times), and their respective rankings

team	run	MAE	RANK(MAE)	RMSE	RANK(RMSE)
fcool	locf	0.20 ± 0.20	1	0.49 ± 0.20	1
idppexplorers	naive	0.20 ± 0.22	1	0.49 ± 0.22	1
unipd	hold	0.20 ± 0.21	1	0.49 ± 0.21	1
mandatory	d1	0.20 ± 0.19	1	0.49 ± 0.19	1
idppexplorers	EN	0.22 ± 0.17	2	0.50 ± 0.17	2
CBMUnito	RF-MonoWindow	0.23 ± 0.19	3	0.52 ± 0.19	3
bitua	ensemble-max	0.25 ± 0.18	4	0.54 ± 0.18	4
bitua	temporalAnalysis	0.29 ± 0.24	5	0.61 ± 0.24	6
unipd	average	0.33 ± 0.18	6	0.60 ± 0.18	5
unipd	logistic-ALSFRS	0.34 ± 0.21	7	0.64 ± 0.21	8
fcool	RFClassifier	0.35 ± 0.22	8	0.68 ± 0.22	15
unipd	rf	0.36 ± 0.22	9	0.65 ± 0.22	11
idppexplorers	voting	0.37 ± 0.15	10	0.65 ± 0.15	10
bitua	moremetrics	0.37 ± 0.23	10	0.68 ± 0.23	16
mandatory	12hist14	0.37 ± 0.19	11	0.65 ± 0.19	9
unipd	rf-reg	0.37 ± 0.19	12	0.64 ± 0.19	7
mandatory	1hist09	0.38 ± 0.31	13	0.72 ± 0.31	30
bitua	median	0.38 ± 0.23	14	0.70 ± 0.23	20
fcool	2nd-best-both-metrics	0.39 ± 0.26	15	0.71 ± 0.26	25
bitua	mean	0.39 ± 0.26	15	0.71 ± 0.26	21
mandatory	1hist05	0.39 ± 0.20	16	0.66 ± 0.20	12
unipd	ridge	0.39 ± 0.20	17	0.69 ± 0.20	17
idppexplorers	gb	0.40 ± 0.18	18	0.69 ± 0.18	18
mandatory	1hist04	0.40 ± 0.26	18	0.66 ± 0.26	13
mandatory	12hist10	0.41 ± 0.23	19	0.67 ± 0.23	14
unipd	optrun	0.41 ± 0.19	20	0.71 ± 0.19	22
idppexplorers	svm	0.41 ± 0.23	20	0.75 ± 0.23	33
fcool	best-both-metrics	0.41 ± 0.22	20	0.71 ± 0.22	26
mandatory	12hist13	0.42 ± 0.24	21	0.72 ± 0.24	28
bitua	ensemble-avg	0.42 ± 0.23	22	0.71 ± 0.23	24
idppexplorers	lr	0.42 ± 0.20	23	0.73 ± 0.20	32
mandatory	1hist03	0.42 ± 0.24	24	0.69 ± 0.24	19
mandatory	12hist11	0.43 ± 0.28	25	0.72 ± 0.28	27
fcool	3rd-best-both-metrics	0.43 ± 0.26	25	0.78 ± 0.26	39
mandatory	d0	0.44 ± 0.14	26	0.72 ± 0.14	29
mandatory	1hist08	0.44 ± 0.26	27	0.71 ± 0.26	23
idppexplorers	et	0.44 ± 0.24	27	0.78 ± 0.24	36
idppexplorers	dt	0.44 ± 0.22	28	0.72 ± 0.22	31
idppexplorers	knn	0.46 ± 0.19	29	0.77 ± 0.19	35
bitua	ensemble-min	0.47 ± 0.30	30	0.80 ± 0.30	40
idppexplorers	bestModels	0.47 ± 0.24	31	0.81 ± 0.24	42
idppexplorers	lstm	0.48 ± 0.27	32	0.82 ± 0.27	43
mandatory	1hist07	0.48 ± 0.21	33	0.75 ± 0.21	34
mandatory	1hist02	0.48 ± 0.32	34	0.78 ± 0.32	37
idppexplorers	nn	0.49 ± 0.24	35	0.80 ± 0.24	41
mandatory	1hist06	0.49 ± 0.29	36	0.78 ± 0.29	38
idppexplorers	rf	0.51 ± 0.29	37	0.86 ± 0.29	47
fcool	LogisticRegression	0.51 ± 0.28	38	0.84 ± 0.28	46
idppexplorers	bagging	0.51 ± 0.35	39	0.89 ± 0.35	49
unipd	logistic	0.51 ± 0.27	40	0.83 ± 0.27	45
fcool	SVC	0.54 ± 0.34	41	0.89 ± 0.34	48
fcool	XGBClassifier	0.57 ± 0.15	42	0.83 ± 0.15	44
fcool	majority-class	0.66 ± 0.52	43	1.09 ± 0.52	50

which show more stability in the challenge dataset, and where sensor data might not be as effective in detecting eventual changes.

Four teams - iDPPExplorers, Mandatory, FCOOL, and UNIPD - employed this strategy in one of their runs for Task 1, achieving the lowest errors with both metrics (0.20 MAE and 0.49 RMSE) and securing joint first place. The full error scores and rankings for all submitted runs are reported in Table 4.

Note that other runs, which also utilize sensor data, demonstrate performance very close to the first place. Due to the small size of the test set, error estimates exhibit large standard deviations, making it impossible to assert significant differences in the top scores.

7.2 Task 2: Predicting Patient Self-assessment Score from Sensor Data (ALS)

Task 2 is very similar to Task 1, with several teams employing the same methods as they did for Task 1. However, in Task 2, the ALSFRS-R assessments by patients are less regular in timing and less consistent in scoring compared to assessments by clinicians, although they are generally more closely spaced.

The predict-the-last-scores approach remains the top performer, albeit with slightly higher errors (0.29 MAE and 0.58 RMSE), placing the UNIPD and FCOOL teams in joint first place again. Full results are reported in Table 5.

7.3 Task 3: Predicting Relapses from EDSS Sub-scores and Environmental Data (MS)

Table 6 displays the RMSE and MAE scores for all submissions made for Task 3, with consistent scoring positions across both metrics. Additionally, the scatter plot in Fig. 1 offers a visual representation of the performance of all submitted runs, where the x-axis denotes actual values and the y-axis represents predicted values. Ideally, perfect predictions would result in points aligning along a straight line with a slope of 1.

The top-performing strategy is associated with the `UNIPV_t3_rf` run [3], which employs a Random Forest (RF) model after thorough preprocessing stages. Regarding the adoption of environmental features, it is notable that all submissions from the UNIPV (Stefagroup) incorporate environmental variables for relapse predictions. In contrast, the UNIPD team offers both methods with and without the inclusion of environmental variables, achieving their best results with the `UNIPD_t3_ridge_noenv` run, which excludes environmental variables [10].

7.4 Approaches

In this section, we provide a short summary of the approaches adopted by participants in iDPP@CLEF. There are two separate sub-sections, one for Task 1 and 2 – focused on ALS progression prediction – and one for Task 3 – which concerns the MS relapse prediction, using environmental data.

Table 5. For both MAE and RMSE, results are reported as the average error across all twelve ALSFRS-R scores, followed by their average standard deviation (computed by bootstrapping the test set one thousand times), and their respective rankings

team	metric run	MAE	RANK(MAE)	RMSE	RANK(RMSE)
fcool	locf	0.29 ± 0.15	1	0.58 ± 0.15	1
unipd	hold	0.29 ± 0.15	1	0.58 ± 0.15	1
CBMUnito	RF-MonoWindow	0.31 ± 0.16	2	0.60 ± 0.16	2
bitua	ensemble-max	0.33 ± 0.14	3	0.61 ± 0.14	3
bitua	moremetrics	0.37 ± 0.17	4	0.65 ± 0.17	4
bitua	mean	0.39 ± 0.18	5	0.71 ± 0.18	8
bitua	median	0.40 ± 0.21	6	0.69 ± 0.21	5
fcool	2nd-best-both-metrics	0.41 ± 0.15	7	0.71 ± 0.15	6
bitua	ensemble-avg	0.42 ± 0.22	8	0.71 ± 0.22	7
bitua	idpp2024-bitua	0.43 ± 0.24	9	0.72 ± 0.24	9
unipd	average	0.49 ± 0.20	10	0.78 ± 0.20	12
fcool	3rd-best-both-metrics	0.50 ± 0.13	11	0.78 ± 0.13	10
unipd	logistic-ALSFRS	0.50 ± 0.19	11	0.85 ± 0.19	18
bitua	ensemble-min	0.50 ± 0.24	12	0.82 ± 0.24	14
unipd	rf	0.52 ± 0.20	13	0.78 ± 0.20	11
unipd	rf-reg	0.52 ± 0.12	14	0.82 ± 0.12	13
fcool	best-both-metrics	0.53 ± 0.20	15	0.84 ± 0.20	15
fcool	RFClassifier	0.53 ± 0.24	16	0.85 ± 0.24	17
unipd	ridge	0.55 ± 0.27	17	0.85 ± 0.27	16
fcool	LogisticRegression	0.57 ± 0.21	18	0.89 ± 0.21	19
fcool	XGBClassifier	0.59 ± 0.17	19	0.93 ± 0.17	20
unipd	optrun	0.61 ± 0.27	20	0.96 ± 0.27	21
unipd	logistic	0.66 ± 0.29	21	0.99 ± 0.29	22
fcool	SVC	0.67 ± 0.19	22	1.01 ± 0.19	23
ubcs	features100	0.82 ± 0.43	23	1.20 ± 0.43	26
ubcs	featuresall	0.89 ± 0.41	24	1.25 ± 0.41	27
ubcs	features10	0.94 ± 0.49	25	1.33 ± 0.49	28
ubcs	features25	0.96 ± 0.21	26	1.14 ± 0.21	24
ubcs	features20	1.02 ± 0.24	27	1.18 ± 0.24	25
fcool	majority-class	1.03 ± 0.44	28	1.47 ± 0.44	29
ubcs	features50	1.11 ± 0.51	29	1.51 ± 0.51	30

Tasks 1 and 2. Silva and Oliveira [14] (Team BIT.UA) focus on Tasks 1 and 2. Their proposed approaches employ machine learning techniques that rely on RF ensembles. They observed that the most effective solutions are based on temporal analysis, with the maximization strategy being the top-performing approach. Additionally, they emphasize the importance of proper handling of missing data. The authors noted inconsistent performance across the two tasks. Specifically, their approaches tended to be more effective on Task 1, while performance on

Table 6. RMSE and MAE results for all the submitted runs for Task 3

Team	RMSE	MAE
UNIPD_t3_ridge	89.84	68.59
UNIPD_t3_rf_reg	79.74	66.63
UNIPD_t3_average	79.26	65.80
UNIPD_t3_ridge_noenv	78.62	61.37
UNIPV_t3_lmer_last	72.51	47.74
UNIPV_t3_lmer_first	48.07	28.05
UNIPV_t3_rf	41.52	22.49

Task 2 was less satisfactory. Silva and Oliveira attribute this behavior to the variability of the underlying data: Task 1 data, produced by clinicians, was more stable, whereas Task 2 data, produced directly by patients, appeared to be less stable.

Barducci et al. [1] (Team CompBiomedUniTO) tested different approaches to preselect the sensor features to be fed to a RF Classifier. The first solution exploits the mono window approach, which keeps only sensor data recorded within 7 days before the considered questionnaire. The other approach instead considers two windows: the first window is the same as before, and the second window instead considers sensor data recorded when the previously available questionnaire occurred. The second approach aims to provide the model with more information about the changes over time. However, the irregularity of sensor data penalizes the two-windows approach. Indeed, 20 out of 54 patients did not have two 7-day periods with a minimum of 3 days of sensor data. As a result, only the model using the mono window approach was submitted. In general, the results vary significantly depending on the questionnaire and showed better performance for the first task. The lower error in Task 1 may be due to the questionnaire being completed by clinical staff, whose responses are typically more reliable and objective compared to the subjective opinions provided by patients. To address the raised issue, data augmentation is proposed as a possibile solution to increase the number of questionnaires in the training set. In this way, deep learning models could be tested to improve predictions and leverage longer sensor data sequences.

Martins et al. [11] (Team FCOOL) proposed a methodology consisting of independent multi-class models, each predicting a distinct ALSFRS-R question. The authors tested four classification models: Logistic Regression, RF, XGBoost, and Support Vector Machine. To manage sensor data, they first derived static features from the longitudinal data via summarization techniques, and then reduced the feature set using three methods: top-k selection across questions, top-k selection by question, and biclustering. In both tasks, RF achieved top performance among the considered models, but failed to outperform the Last-Observation Carried Forward (LOCF) baseline, except for a few individual questions. Moreover, no consensus was found about the best feature selection or

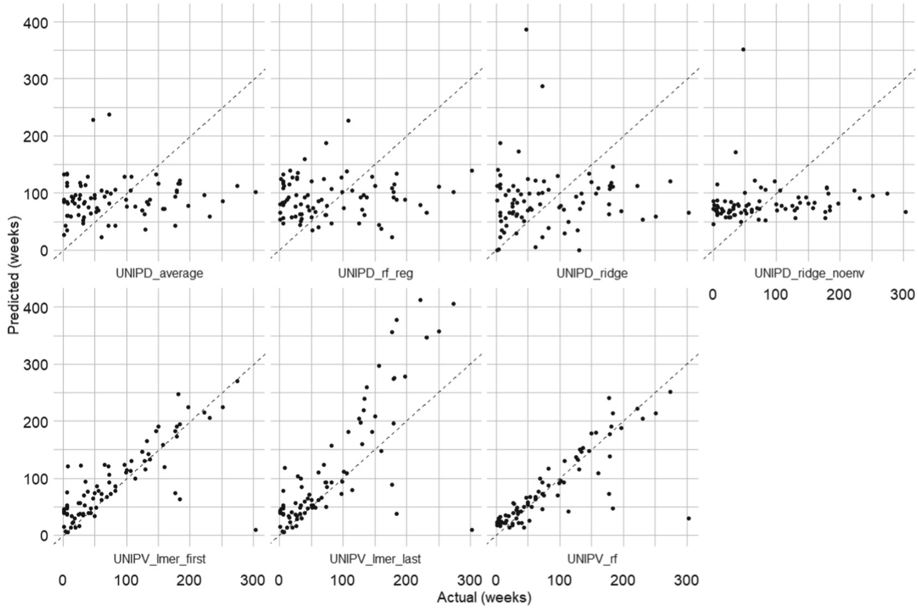


Fig. 1. Actual versus Predicted values for each run submitted for Task 3

extraction approach. Instead, top-k selection by question was the best approach in Task 1, while biclustering in Task 2.

Mehta et al. [12] (Team iDPPExplorers) submitted runs only for Task 1 but analyze the approaches for Task 2 on their working notes paper. Their work focuses on handling the temporal aspect of the sensor data, by studying how to compress it via statistical methods that provide interpretability. Among the set of approaches tested in their work, Mehta et al. observe that the optimal performance is achieved by both a naive baseline and ElasticNet regression. Nevertheless, the authors also observe that, despite the similar performance, the ElasticNet model is more robust and allows a better understanding of the contribution of various features. While they did not take part in Task 2, they observed that the proposed approach is able to achieve better results on self-assessed data provided by the patients. Finally, their conclusive remark hints that, while this preliminary analysis did not highlight any major benefit of using sensor data, a larger dataset with a more diverse set of patients might lead to different conclusions.

In Tasks 1 and 2, Martinello et al. [10] (Team UNIPD) developed a broad set of predictive models based on different methodological approaches using different subsets of the provided variables. The aim of their study was to evaluate whether considering wearable data to predict ALS disability leads to better performance with respect to models that only consider disease-specific variables collected during routine visits. They observe that collecting data from wearable devices can improve the prediction of ALS disability status. However, patients must be properly trained to use the sensors correctly in order to acquire high-quality data leading to significant datasets. Otherwise, if the quality of the acquired wearable data is poor, predicting the next visit ALSFRS-R score by simply holding the current one seems to be a better approach. This is especially true when predicting scores that are self-assigned by patients (Task 2), who seem to be more stable and conservative with respect to their clinician during the disability evaluation process over time.

Okere et al. [13] (Team UBCS) explores different deep-learning techniques to process data, especially to handle missing values. In particular, the authors exploit auto-encoders and multiple imputation techniques to handle missing values and use a RF algorithm to select relevant features. Subsequently, four deep neural networks, such as Multi-Layer Perceptron (MLP), Feed Forward Neural Network (FFNN), Recurrent Neural Network (RNN), and Long-Short Term Memory (LSTM), were trained to perform the two tasks. Experimental results revealed that ensemble predictive models, such as the XGBoost algorithm, show better performance than deep learning models. The authors link the low performance of the models with the small size of the training data.

Task 3. Bosoni et al. [3] (Team Stefagroup) used Topological Data Analysis to compute personal exposure patterns and then employed two predictive approaches. The former relied on applying Linear Regression, RF, and XGBoost to the last follow-up data. The latter used Mixed-Effects modeling on longitudinal data from first to last follow-up. The results showed that incorporating environmental variables provides information statistically significant for predicting relapses. This outcome underlined the need for better methods to compute personal pollution exposure patterns, thereby enhancing the precision of MS progression predictions.

In Task 3 Martinello et al. [10] (Team UNIPD) developed a broad set of predictive models based on different methodological approaches using different subsets of the provided variables. The aim of their study was to evaluate whether considering environmental data to predict MS relapses leads to better performance with respect to models that only consider disease-specific variables collected during routine visits. They observe that environmental data can be beneficial for predicting the occurrence of MS relapses, however, better solutions should be explored to refine the data collection and variable extraction process in order to obtain more precise and focused predictions.

8 Conclusions and Future Work

iDPP@CLEF 2024 is the third and last iteration of the iDPP@CLEF evaluation campaign. The focus of this evaluation campaign was on developing AI models capable of preemptively estimating the risks that patients affected by ALS and MS will need medical support and to describe the progression of their disease, to foster patient stratification and aid clinicians in providing the due care in the most effective and rapid way.

iDPP@CLEF 2024 operated in continuation with iDPP@CLEF 2022 and iDPP@CLEF 2023, expanding previously proposed tasks, but also identifying novel tasks. In particular, iDPP@CLEF was organized into three tasks. The first two tasks focused on predicting the ALSFRS-R for patients affected by ALS, using data collected via environmental sensors and wearable devices. This makes iDPP@CLEF 2024 the first edition of making use of data collected on patients currently involved in the BRAINTEASER project. The third task of iDPP@CLEF 2024 built upon the results of iDPP@CLEF 2023, by focusing on the prediction of the disease progression of patients affected by MS. More in detail, this task focused on predicting when an MS patient will experience a relapse. As an improvement over the previous iDPP edition, this year participants were also provided with environmental data that could have been used to improve the AI models.

In terms of participation, 28 teams registered in the Lab, suggesting overall interest in the topic from the research community, and 8 teams were able to submit their results for a total of 97 submitted runs. The task that received the most interest was the first, with 59 submissions alone.

While this cycle concludes the evaluation campaign of iDPP@CLEF, we envision several possible research paths for which iDPP@CLEF paved the way. First of all, novel and more effective AI approaches can be developed in the future, by using iDPP@CLEF data as training and evaluation sets. Secondly, iDPP@CLEF has identified several guidelines and good practices that can be adapted to devise novel shared tasks and evaluation campaigns in the future, either concerning ALS and MS, other neurological diseases, or the medical domain at large.

Acknowledgments. The work reported in this paper has been partially supported by the BRAINTEASER (https://brainteaser.health/) project (contract n. GA101017598), as a part of the European Union's Horizon 2020 research and innovation programme.

References

1. Barducci, G., Sartori, F., Birolo, G., Sanavia, T., Fariselli, P.: ALSFRS-R score prediction for amyotrophic lateral sclerosis. In: CLEF 2024 Working Notes (2024)
2. Birolo, G., et al.: Overview of iDPP@CLEF 2024: the intelligent disease progression prediction challenge. In: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9th to 12th September 2024
3. Bosoni, P., et al.: Predicting multiple sclerosis relapses using patient exposure trajectories. In: CLEF 2024 Working Notes (2024)
4. Cedarbaum, J.M., et al.: The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. J. Neurol. Sci. **169**(1–2), 13–21 (1999)
5. Faggioli, G., et al.: Intelligent disease progression prediction: overview of IDPP@CLEF 2023. In: Arampatzis, A., et al. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction - 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18-21, 2023, Proceedings, LNCS, vol. 14163, pp. 343–369, Springer, Cham (2023). https://doi.org/10.1007/978-3-031-42448-9_24

6. Faggioli, G., et al.: Overview of IDPP@CLEF 2023: the intelligent disease progression prediction challenge. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, 18th to 21st September 2023, CEUR Workshop Proceedings, vol. 3497, pp. 1123–1164, CEUR-WS.org (2023). https://ceur-ws.org/Vol-3497/paper-095.pdf
7. Guazzo, A., et al.: Intelligent disease progression prediction: overview of iDPP@CLEF 2022. In: Barrón-Cedeño, A., et al. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022), LNCS, vol. 13390, pp. 395–422. Springer, Heidelberg, Germany (2022). https://doi.org/10.1007/978-3-031-13643-6_25
8. Guazzo, A., et al.: Overview of iDPP@CLEF 2022: the intelligent disease progression prediction challenge. In: Faggioli, G., Ferro, N., Hanbury, A., Potthast, M. (eds.) CLEF 2022 Working Notes, pp. 1130–1210, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073 (2022). http://ceur-ws.org/Vol-3180/
9. Küffner, R., et al.: Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. Nat. Biotechnol. **33**(1), 51–57 (2015)
10. Marinello, E., et al.: Using wearable and environmental data to improve the prediction of amyotrophic lateral sclerosis and multiple sclerosis progression: an explorative study. In: CLEF 2024 Working Notes (2024)
11. Martins, A., et al.: Predicting the functional rating scale and self-assessment status of ALS patients with sensor data. In: CLEF 2024 Working Notes (2024)
12. Mehta, R., Pramov, A., Verma, S.: Machine learning for ALSFRS-R score prediction: making sense of the sensor data. In: CLEF 2024 Working Notes (2024)
13. Okere, C., Thuma, E., Mosweunyane, G.: UBCS at IDPP: predicting patient self-assessment score from sensor data using machine learning algorithms. In: CLEF 2024 Working Notes (2024)
14. Silva, J., Oliveira, J.: BIT.UA at IDPP: predictive analytics on ALS disease progression using sensor data with machine learning. In: CLEF 2024 Working Notes (2024)
15. World Health Organization: WHO global air quality guidelines: Particulate matter (PM(2.5) and PM(10)), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. World Health Organization, Geneva (2021). ISBN 978-92-4-003422-8, review

Overview of the ImageCLEF 2024: Multimedia Retrieval in Medical Applications

Bogdan Ionescu[1(✉)], Henning Müller[2], Ana-Maria Drăgulinescu[1], Johannes Rückert[3], Asma Ben Abacha[4], Alba García Seco de Herrera[5,6], Louise Bloch[3], Raphael Brüngel[3], Ahmad Idrissi-Yaghir[3], Henning Schäfer[8], Cynthia Sabrina Schmidt[7], Tabea M. G. Pakull[8], Hendrik Damm[3], Benjamin Bracke[3], Christoph M. Friedrich[3], Alexandra-Georgiana Andrei[1], Yuri Prokopchuk[9], Dzmitry Karpenka[9], Ahmedkhan Radzhabov[9], Vassili Kovalev[9,10], Cécile Macaire[11], Didier Schwab[11], Benjamin Lecouteux[11], Emmanuelle Esperança-Rodier[11], Wen-Wai Yim[4], Yujuan Fu[12], Zhaoyi Sun[12], Meliha Yetisgen[12], Fei Xia[11], Steven A. Hicks[13], Michael A. Riegler[13], Vajira Thambawita[13], Andrea Storås[13], Pål Halvorsen[13], Maximilian Heinrich[14], Johannes Kiesel[14], Martin Potthast[15,16,17], and Benno Stein[14]

[1] National University of Science and Technology Politehnica Bucharest, Bucharest, Romania
bogdan.ionescu@upb.ro
[2] University of Applied Sciences Western Switzerland (HES-SO), Delémont, Switzerland
[3] Department of Computer Science, University of Applied Sciences and Arts, Dortmund, Germany
[4] Microsoft, Redmond, USA
[5] University of Essex, Colchester, UK
[6] UNED, Madrid, Spain
[7] Institute for Artificial Intelligence in Medicine, University Hospital Essen, Essen, Germany
[8] Institute for Transfusion Medicine, University Hospital Essen, Essen, Germany
[9] Belarusian National Academy of Sciences, Minsk, Belarus
[10] Belarus State University, Minsk, Belarus
[11] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
[12] University of Washington, Seattle, USA
[13] SimulaMet, Oslo, Norway
[14] Bauhaus-Universität Weimar, Weimar, Germany
[15] University of Kassel, Kassel, Germany
[16] hessian.AI, Darmstadt, Germany
[17] ScaDS.AI, Leipzig, Germany

Abstract. This paper presents an overview of the ImageCLEF 2024 lab, organized as part of the Conference and Labs of the Evaluation Forum – CLEF Labs 2024. ImageCLEF, an ongoing evaluation event since 2003, encourages the evaluation of technologies for annotation, indexing and

L. Goeuriot et al. (Eds.): CLEF 2024, LNCS 14959, pp. 140–164, 2024.
https://doi.org/10.1007/978-3-031-71908-0_7

retrieval of multimodal data. The goal is to provide information access to large collections of data across various usage scenarios and domains. In 2024, the 22st edition of ImageCLEF runs three main tasks: (i) a *medical task*, continuing the caption analysis, Visual Question Answering for colonoscopy images alongside GANs for medical images, and medical dialogue summarization; (ii) a novel task related to *image retrieval/generation for arguments* for visual communication, aimed at augmenting the effectiveness of arguments; and (iii)ToPicto, a new task focused on *translating natural language*, whether spoken or textual, into a sequence of pictograms. The benchmarking capaign was a real success and received the participation of over 35 groups submitting more than 220 runs.

Keywords: Medical text summarization · medical image caption analysis · visual question answering · Generative Adversarial Networks (GANs) · image retrieval · translation of neural language · ImageCLEF lab

1 Introduction

This paper presents the overview of the ImageCLEF 2024 lab, part of the Conference and Evaluation Forum - CLEF Labs 2024. Started in 2003, ImageCLEF[1] is an ongoing evaluation initiative that promotes the evaluation of technologies for annotation, indexing, and retrieval of visual data, facilitating information access to image collections across diverse domains. Over the years, ImageCLEF has continually adapted to emerging trends, adding tasks ranging from general object classification and retrieval to specialized application areas such as medical imaging, social media, nature, and security.

Over the years, ImageCLEF and also CLEF have shown a strong scholarly impact that was assessed in [45,46]. For instance, the term "ImageCLEF" returns on Google Scholar[2] over 7,390 article results (search on June 11, 2024). This underlines the importance of the evaluation campaigns for disseminating best scientific practices.

In 2024, the 24th edition of ImageCLEF features three main tasks: i) a medical task continuing the four sub-tasks from the previous edition [24] (the 8th edition of the Caption task, the 5th edition of the MEDIQA task, and the 2nd editions for GANs and MedVQA tasks), ii) ToPicto, a new task focusing on augmentative and alternative communication using pictograms, and iii) Image Retrieval for Arguments, a new task for ImageCLEF lab, organized in collaboration with Touché lab.

2 Overview of Tasks and Participation

ImageCLEF 2024 [23] consists of three main tasks to cover a *diverse range* of multimedia retrieval in *medical applications*. It followed the 2019 tradition [25]

[1] http://www.imageclef.org/.
[2] https://scholar.google.com/.

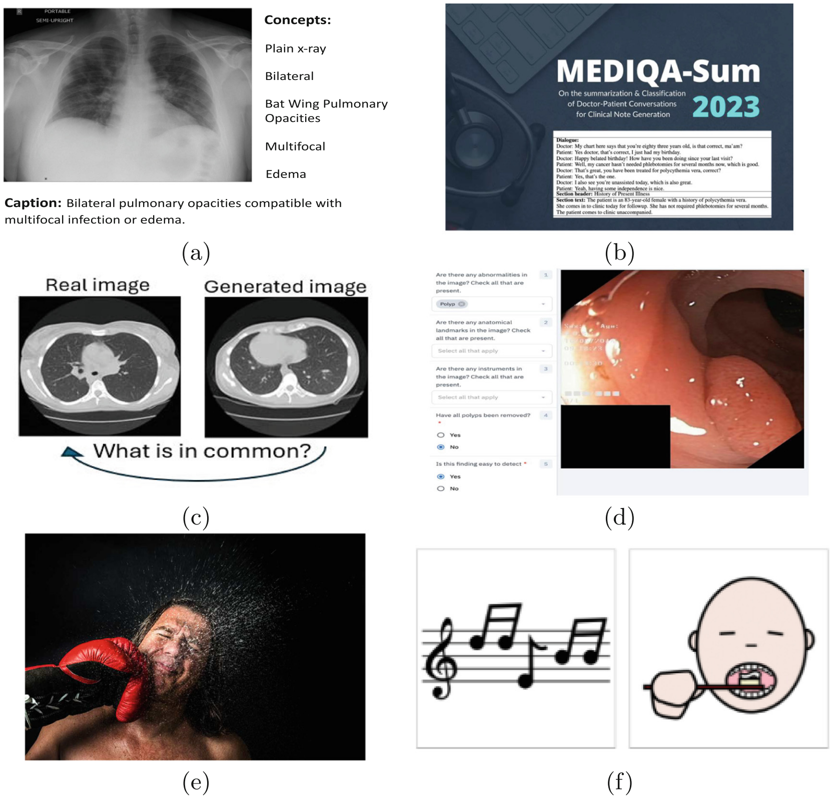


Fig. 1. Sample images: (a) ImageCLEFmedical-caption with an image and the corresponding CUIs and captions, (b ImageCLEFmedical-MEDIQA-MAGIC with an example of doctor-patient conversation, (c) ImageCLEFmedical-GAN with an example of real and generated images, (d) ImageCLEFmedical-VQA with examples of questions and answers in the area of colonoscopy, (e)) Argument-Image with a picture of a boxer conveying the premise that boxing causes injuries (see Footnote 3), (f) ToPicto from left to right: “music”, “brush the teeth”.

of diversifying the use cases [3,20,35,41,44,51]. The 2024 tasks are presented as follows:

- **ImageCLEFmedical**. Since 2004, the ImageCLEF benchmarking initiative has included medical tasks. However, by 2018, although nearly all tasks were medical, there was minimal interaction between them. Therefore, beginning in 2019, the medical tasks were consolidated into a single task centered around a specific problem, with multiple subtasks. This approach fostered synergies between the different domains:
  - *Caption*: This is the 8th edition of the task in this format, however, it is based on previous medical tasks. The task is once again running with both the “concept detection” and “caption prediction” subtasks [40], after the

former was brought back in 2021 due to participants' demands [18,34,41]. The "caption prediction" subtask focuses on composing coherent captions for the entirety of a radiology image, while the "concept detection" subtask focuses on identifying the presence of relevant concepts in the same corpus of radiology images. After a smaller data set of manually annotated radiology images was used in 2021, the 2024 edition once again uses a larger dataset based on Radiology Objects in COntext version 2 (ROCOv2) [42], which was already used in 2019–2023.[3]

- *MEDIQA-MAGIC*: This is the fifth edition of the MEDIQA tasks and its second edition in the text format. The 2019 MEDIQA task featured several medical natural language semantic retrieval-related tasks, including natural language inference (NLI) classification of MIMIC-III clinical note sentences, recognizing question entailment (RQE) in consumer health questions, and reranking retrieved answers to consumer health questions. Continuing in 2021, the next MEDIQA task resumed hosting one clinical subtask and two consumer-health question-answer related subtasks [7]. Different from the 2019 subtasks, MEDIQA 2021 focused on summarization; summarization of clinical radiology note findings, consumer health questions, and consumer health answers. 2023 edition included clinical dialogue section header classification, short-dialogue note summarization, and full-encounter generation. The MEDIQA-MAGIC 2024 task mirrors the setup of the MEDIQA-M3G task. Participants receive a consumer health textual query along with associated images and are tasked with producing a preliminary doctor response. Responses are evaluated against two reference standards using deltaBLEU [17], BERT-score [53], and UMLS-F1 (F1 score of UMLS concept combined with an assertion label).
- *GANs*: In this edition, we continue to study the first sub-task illustrated in Fig. 1 – "Detect generative models" "fingerprints"– proposed in the previous edition [3] focused on examining the existing hypothesis that GANs generate medical images containing certain "fingerprints" of the authentic images used for generative network training. We extended the task by investigating this hypothesis for two different generative models. Another sub-task is introduced to this second edition—Detect generative models' "fingerprints". The second sub-task explores the hypothesis that generative models imprint unique fingerprints on generated images and whether different generative models or architectures leave discernible signatures within the synthetic images they produce [4].
- *MEDVQA-GI*: The MEDVQA-GI challenge is held for the second time this year with a new goal. One of the new frontiers in AI-driven medical diagnosis is the application of text-to-image generative models. This area integrates language processing and image synthesis to enhance diagnostic capability in the medical field. In this task, we aim to direct the power of artificial intelligence to generate medical images based on text input, along with optimal prompts for off-the-shelf generative models building

[3] Source: Sweating fighter is punched in the face - gettyimages.

up on the dataset collected in the first edition of MEDVQA-GI. The objective is to improve the diagnosis and classification of real medical images using AI-generated imagery. The task is divided into two main subtasks

- **Image Retrieval/Generation for Arguments** (Argument-Image). This is the third edition of the task. Pictures are a powerful means of visual communication and can be used to enhance the impact of arguments. This observation leads to our task where, given an argument, participants shall find images that help to convey the argument's premise. In this context "convey" is meant in a general manner; it can depict what is described in the argument, but it can also show a generalization (e.g., a symbolic image that illustrates a related abstract concept) or specialization (e.g., a concrete example). To better explain why an image conveys a premise, participants can optionally submit a rationale that helps explain why an image is relevant. This is a joint task with Touché 2024. Details on this task are provided in the Touché overview paper [27]. In Fig. 1 we see an example submission for an argument, which consists of the premise "Boxing can lead to serious injuries". and the claim "Boxing is a dangerous sport!"
- **ToPicto**. This is the first edition of the task. The objective of ToPicto is to investigate the translation of natural language, either speech or text, into a sequence of pictograms as depicted in Fig. 1. Generating pictograms is an emerging and significant domain in natural language processing, with multiple potential applications. It can enable communication with individuals who have disabilities, aid in medical settings for individuals who do not speak the language of a country, and also enhance user understanding in the service industry. Recent advances in artificial intelligence and machine translation have greatly improved performance in text-to-text as well as speech-to-text translations, but they have not been applied to text-to-pictogram and speech-to-pictogram translations before. ImageCLEFtoPicto seeks to bring together linguists, computer scientists, and translators to develop new translation methods. ImageCLEFtoPicto is divided into two subtasks:
  - *Text-to-Picto*: The first proposed subtask focuses on the automatic generation of a corresponding sequence of pictogram terms from a French text.
  - *Speech-to-Picto*: Building on the first subtask, Speech-to-Picto focuses on the two modalities speech and pictograms. The objective is to directly translate speech to a sequence of pictograms without going through the transcription dimension, which is the focus of the speech community with current spoken language translation systems.

In order to participate in the evaluation campaign, research groups were required to register by following the instructions on the ImageCLEF 2024 webpage[4]. This year, the challenges were organized through the AI4Media bench-

[4] https://www.imageclef.org/2024/.

Table 1. Key figures regarding participation in ImageCLEF 2024.

Task	Groups that submitted results	Submitted runs	Submitted working notes
Caption	14	82	13
MEDIQA-MAGIC	3	22	3
GANs	11	100	10
MedVQA	2	6	2
Argument-Image	2	8	2
ToPicto	4	7	4
Overall	33	225	34

marking platform[5] based on codalab[6]. Similar to previous editions, participants were required to submit a signed End User Agreement (EUA) to access the datasets. Table 1 summarizes the participation in ImageCLEF 2024, indicated the statistics both per task and for the overall lab. The table also shows the number of groups that submitted runs and the ones that submitted a working notes paper describing the techniques used. Teams were allowed to register for several tasks. Following a decline in participation in 2016, there was an increase in 2017, 2018 and 2019. Specifically, in 2018, 31 teams completed the tasks and 28 working notes papers were submitted. In 2019, the number of participating teams climbed to 63, and we received 50 working papers. In 2020, 40 teams completed the tasks and submitted their working notes papers. In 2022, participation decreases with 28 teams completing the tasks and 26 working notes paper submitted. There was a new increase in 2023 with 47 teams submitting results and 39 working notes papers received. This year's edition of ImageCLEF attracted 36 teams and we received 34 working notes. The number of runs dropped compared to 2022 and 2023 with more teams involved 256 (2022) and 241 (2023) vs 225 (2024). This could be due to teams focusing on developing higher-quality solution and the increased complexity of the tasks this year, which may have required more time and resources per run.

In the following sections, we present the tasks. Only a short overview is reported, including general objectives, description of the tasks and data sets, and a short summary of the results. A detailed review of the received submissions for each task is provided with the task overview working notes: Caption [40], Mediqa [50], GAN [4], MedVQA [21], ToPicto [29] and Image Retrieval for Arguments [27].

3 The Caption Task

The caption task was first proposed as part of the ImageCLEFmedical [18] in 2016, aiming to extract the most relevant information from medical images.

[5] https://ai4media-bench.aimultimedialab.ro/.
[6] https://github.com/AIMultimediaLab/Ai4media-Bench.

Hence, the task was created to condense visual information into textual descriptions. With the exception of 2019 and 2020, when only the concept detection task was offered, the ImageCLEFmedical Caption task has been running since 2017 with two subtasks: concept detection and caption prediction. With a break in 2021, where fewer images which were all manually annotated by medical doctors were used, an extended version of the ROCO data was set was used from 2019 to 2023 [41] for both subtasks, while the 2023 edition switched from BLEU-1 [32] to BERTScore [54] as the primary evaluation metric for caption prediction. In the 2024 edition of the ImageCLEFmedical Caption [40], the data used for both subtasks was based on the newly released ROCOv2 [42] data set.

3.1 Task Setup

The ImageCLEFmedical 2024 Caption [40] follows the format of the previous ImageCLEFmedical Caption tasks. In 2024, the overall task comprises two subtasks: "Concept Detection" and "Caption Prediction". The concept detection sub-task focuses on predicting Unified Medical Language System® (UMLS) Concept Unique Identifiers (CUIs) [12] based on the visual image representation in a given image. The caption prediction subtask focuses on composing coherent captions for the entirety of the images. This year, a new optional, experimental explainability extension has been introduced for the caption prediction task. This extension aims to improve the understanding of the models by asking participants to provide explanations, such as heat maps or Shapley values, for a selected number of images. These explanations are manually reviewed to assess their effectiveness and clarity.

The detected concepts are evaluated using the balanced precision and recall trade-off in terms of F1-scores, as in previous years. Like last year, a secondary F1-score is computed using a subset of concepts that were manually curated and only contain x-ray anatomy and image modality concepts. Similar to last year, BERTScore was used as the primary metric for the evaluation of the caption prediction subtask. BERTScore evaluates the semantic similarity of the predicted captions. In addition to the BERTScore, a secondary ROUGE score, which measures the overlap of content between the predicted captions and reference captions, was provided. After the submission period ended, a number of additional scores were calculated and published: METEOR [5], CIDEr [48], CLIPScore [19], BLEU and BLEURT [43]. This year, two new metrics, MedBERTScore and ClinicalBLEURT [10], were added. These domain-adapted metrics are designed to better assess the relevance and accuracy of generated text in medical contexts, with the goal of improving the precision of evaluations in this specialized field.

3.2 Dataset

In 2024, an updated version of the ROCO dataset, called ROCOv2 [42], is utilized for both subtasks. The ROCOv2 dataset is derived from biomedical articles of the PMC Open Access Subset[7] [38] and was extended with new images added

[7] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/.

Table 2. Performance of the participating teams in the ImageCLEFmedical 2024 Caption concept detection subtask. The best run per team is selected. Teams with previous participation in 2023 are marked with an asterisk.

Team	F1	Secondary F1
DBS-HHU	0.6375	0.9534
AUEB-NLP-Group*	0.6319	0.9393
DS@BioMed	0.6200	0.9312
SSNMLRGKSR*	0.6001	0.9056
UACH-VisionLab	0.5988	0.9363
MICLabNM	0.5795	0.8835
Kaprov	0.4609	0.7301
VIT_Conceptz	0.1812	0.2647
CS_Morgan*	0.1076	0.2105

since the last time the dataset was updated. For this year, only CC BY and CC BY-NC licensed images are included. From the captions, UMLS® concepts were extracted, and concepts regarding anatomy and image modality were manually validated for all images.

Following this approach new training, validation, and test sets were provided for both tasks:

- *Training set* including 70,108 radiology images and associated captions and concepts.
- *Validation set* including 9972 radiology images and associated captions and concepts.
- *Test set* including 17,237 radiology images.

Table 3. Performance of the participating teams in the ImageCLEFmedical 2024 Caption prediction subtask. The best run per team is selected. Teams with previous participation in 2023 are marked with an asterisk.

Team	BERTScore	ROUGE
PCLmed	0.6299	0.2726
CS_Morgan	0.6281	0.2508
DarkCow	0.6267	0.2452
AUEB-NLP-Group	0.6211	0.2049
2Q2T	0.6178	0.2478
MICLab	0.6128	0.2135
DLNU_CCSE	0.6066	0.2179
Kaprov	0.5964	0.1905
DS@BioMed	0.5794	0.1031
DBS-HHU	0.5769	0.1531
KDE-medical-caption	0.5673	0.1325

3.3 Participating Groups and Submitted Runs

In the eighth edition of the ImageCLEFmedical Caption task, 50 teams registered and signed the End-User-Agreement that is needed to download the development data. 14 teams submitted 82 graded runs (13 teams submitted working notes) attracting similar attention to 2023. Similar to last year, participants did not have access to their own scores until after the submission period was over. Of the 9 teams that participated in the concept detection subtask this year, 4 also participated in 2023. Of the 11 teams which submitted runs to the caption prediction subtask, 6 also participated in 2023. Overall, 6 teams participated in both subtasks, and 5 teams participated only in the caption prediction subtask. Unlike in 2023, 3 teams participated only in the concept detection subtask.

In the concept detection subtasks, the groups used primarily multi-label classification systems.

The winning team this year utilized an ensemble of four different CNNs. In the caption prediction subtask, teams primarily utilized encoder-decoder frameworks with various backbones, including transformer-based decoders and LSTMs [22].

The winning team introduced medical vision-language foundation models (Med-VLFMs) by combining general and specialist vision models to achieve top rankings in the challenge.

To get a better overview of the submitted runs, the primary scores of the best results for each team are presented in Tables 2 and 3.

3.4 Results

For the concept detection subtak, the overall F1 scores increased strongly compared to last year despite very similar approaches being employed by the teams. In addition to continuously improved and scaled-up approaches by the teams, some possible explanations for this include an improved and overall larger dataset, a lower number of unique concepts in the test set, and the removal of directionality concepts.

The same applies for the general view on results of this year's caption prediction task. The top scores were slightly worse for BERTScore, but last year's winners CSIRO did not participate this year. Returning teams improved their scores across the board showing that the dataset for this year is comparable to last year and that while teams have experimented with many different approaches including LLMs for caption generation, no breakthrough improvement has been achieved with these new techniques. The new optional explainability extension was not adpoted by the teams, only the team MICLabNM [14] submitted explainability results after the end of the submission phase.

3.5 Lessons Learned and Next Steps

This year's caption task of ImageCLEFmedical once again ran with both subtasks, concept detection and caption prediction. Like last year, it used a

ROCOv2-based data set for both challenges. Manually validated concepts for X-ray directionality information introduced last year were removed for this year's dataset. It attracted 14 teams who submitted a total of 82 graded runs, a similar level of participation to last year. Changes were introduced in the evaluation metrics, with the addition of two new domain-specific metrics, MedBERTScore and ClinicalBLEURT, specifically for the caption prediction task. These additions were made based on feedback received from participants the previous year.

For next year's ImageCLEFmedical Caption challenge, some possible improvements include an improved caption prediction evaluation metric which is specific to medical texts or a combination of different metrics, as well as additional metrics for readability and factuality. The optional explainability extension might be moved into its own subtask for next year.

4 The MEDIQA-MAGIC Task

Since 2019, the MEDIQA shared-tasks have tackled various question-answering and summarization challenges related to medical reasoning, language, and semantics. Its first edition included the classification tasks of clinical note sentence natural language inference and recognizing question entailment, as well as their application towards answer-retrieval re-ranking. In 2021, the MEDIQA challenges focused on monologue-to-monologue summarization tasks, including clinical radiology note findings summarization, consumer health question summarization, and multiple answer summarization [7]. In 2023, two editions were hosted. Both featured problems related to dialogue-to-monologue summarization for clinical note from doctor-patient conversations. Subtasks included short-dialogue section header and note generation, topic-to-note summarization, full-encounter dialogue-to-note generation, and full-encounter note-to-dialogue generation [8,51]. This year, similarly two related editions were hosted. These tasks revolved around the problem of multi-modal visual question-answering tasks on consumer health dermatology problems. While MEDIQA-M3G [9] (multi-modal, multilingual answer generation), part of the NAACL 2024 ClinicalNLP Workshop focused on short-answers in English, Chinese, and Spanish; the MEDIQA-MAGIC (Multimodal And Generative TelemedICine) task part of ImageCLEF 2024, described here, included in-depth full answer responses for English only.

4.1 Task Setup

The MEDIQA-MAGIC 2024 task follows the setup for the MEDIQA-M3G task. Participants are supplied with a consumer health textual query and associated images. The target objective is to output a draft doctor response. The evaluated responses were graded against two reference standards using deltaBLEU [17], BERTScore [53], and UMLS-F1 (F1 score of UMLS concept combined with an assertion label). For more comprehensive details related to the task, dataset, and results, please refer to the task overview paper [50].

4.2 Dataset

The 2024 MEDIQA-MAGIC challenge used data from the Reddit sub-collection of the DermaVQA dataset [52]. To comply with data usage guidelines, only post id's and our labels were shared with participants. Participants who registered through Reddit could receive API credentials to access Reddit's data. Afterwards, the participants could use supplied download script[8] to retrieve the original input data. As Reddit users may opt to delete content, the final set of test set id's were determined by the subset of test id's retrieval shortly after the submission deadline. The original labeled dataset included 347, 50, 93 instances for train, valid, and test sets. The final number of test set instances was 78.

4.3 Participating Groups and Submitted Runs

Overall 3 teams participated with a total of 22 runs. The teams came from three different countries and continents (India, Poland, and Taiwan).

4.4 Results

The final results are shown in Table 4. The submitted systems represented a variety of solutions, including using out-of-the-box gemini [1] models (YuanAI), applying small visual language models (VisionQAries), and utilizing visual-language encoders with cosine similarities(IRLab@IIT_BHU). The ranges of scores were co-located in the lower spectrum for all three metrics (100 total for BLEU, and 1.0 for BERTScore and UMLS F1), indicating the difficulty of the task.

Table 4. Performance of the participating teams in the MEDIQA-MAGIC 2024 Answer Generation Task (Best Run).

Team	Institution	BLEU	BERTScore	MEDCON
VisionQAries	IIT (BHU), Varanasi, India	8.969	0.844	0.077
IRLab@IIT_BHU	Poland	4.536	0.839	0.066
YuanAI	Yuan Ze University, Taiwan	4.371	0.856	0.087

The following sections briefly describe the teams' solutions. More information can be found in the overview [50]:
IRLab@IIT_BHU manually labeled instances into 160 categories, passing image and text data through CLIP encoders. Text data went through a Bi-LSTM and vision data through an MLP, with their results averaged to create a label vector. Training involved weighted cosine similarity loss. During inference, the combined embedding matched the closest label embedding. They also used data augmentation with TextGenie and GPT2 for classification.

[8] https://github.com/wyim/MEDIQA-MAGIC-2024.

VisionQAries focused on small multi-modal models, testing direct prompting and fine-tuning on moondream2 and TinyLLaVA models. Fine-tuning moondream2 yielded better BLEU scores than direct prompting.
YuanAI used the Gemini image-to-text model, followed by a LoRA fine-tuned Llama3 to process outputs and queries, generating the final response.

4.5 Lessons Learned and Next Steps

This year's ImageCLEF MEDIQA-MAGIC task differed from the 2024 NAACL ClinicalNLP MEDIQA-M3G Shared Tasks [9] by using a different dataset and requiring participants to obtain Reddit credentials, which may have deterred some teams. Another major difference was the longer answer lengths, averaging 90 words compared to 12, increasing the challenge in answer generation and evaluation. The competition used a codabench-based platform for easier submissions and result computation, with an API for automatic participant data download. This year required GitHub code submissions, unlike last year's requirement for run-able code, resulting in less complete documentation. Future editions may use Codabench's real-time inference to ensure clean, run-able code without manual effort.

This task required extensive free-text answers, unlike other visual question-answering tasks with 1–2 word responses, and allowed multiple correct answers, presenting challenges in natural language evaluation. Future editions will address specialty to consumer health multi-modal problems and experiment with evaluation methods for longer text and multiple correct answers.

5 The GANs Task

Biomedical imaging has advanced significantly in recent years due to the convergence of machine learning (ML) and artificial intelligence (AI) technologies, particularly through the development of generative models like Generative Adversarial Networks (GANs). These models have proven effective in producing synthetic images that mimic real biomedical images, creating new opportunities for study and application. Synthetic images produced by these models offer several potential advantages in the biomedical domain, including augmenting existing datasets to address data scarcity and imbalances, which is especially valuable given the difficulty, cost, and time involved in obtaining large amounts of labeled medical data. Moreover, AI algorithms benefit from synthetic images by reducing dependency on real patient data, thus mitigating privacy concerns.

5.1 Task Setup

This is the second edition of the task and consists of two sub-tasks. In addition to the sub-task presented in the previous edition, "Identify training data fingerprints" [3], we have introduced the second sub-task entitled "Detect generative models' fingerprints". The objective of the first sub-task was to detect "fingerprints" within the synthetic biomedical image data to determine which real images were used in training to produce the generated images. The task is formulated as follows:

– *given two sets that contain generated and real images, the participants are requested to employ machine learning and/or deep learning models to determine for each set which of the real images were used to train the model to generate the provided synthetic images.*

The second sub-task explores the hypothesis that generative models imprint unique fingerprints on generated images. The focus is on understanding whether different generative models or architectures leave discernible signatures within the synthetic images they produce. By providing a set of synthetic The task was formulated as follows:

– *given a set of generated images and the number of generative models used, the participants are required to group the images based on the model that generated them.*

5.2 Dataset

The benchmarking image data consists of axial slices of 3D CT images extracted from a bigger dataset of about 8000 lung tuberculosis patients. Considering this, some of the slices may appear pretty "normal" whereas the others may contain certain lung lesions including severe ones. These images are stored as 8-bit/pixel PNG images with dimensions of 256×256 pixels. The artificial slice images are 256×256 pixels in size. The dataset for the first sub-task consisted of both real and generated images, while the dataset for the second sub-task consisted in synthetic images only generated using different generative models.

5.3 Participating Groups and Submitted Runs

Overall, 23 teams registered to both tasks. Among them, 10 teams completed the first sub-task and submitted their runs, while 7 teams completed the second sub-task (including the task organizing team). Notably, 6 teams were common to both sub-tasks, demonstrating consistency across the tasks. When it comes to submitting the working notes, one team did not submit them, resulting in an adherence rate of 90.90%.

5.4 Results

For the first sub-task, "Identify training data fingerprints", a variety of methods were employed, ranging from advanced image preprocessing techniques to deep learning models. Various techniques such as binarization, histogram equalization, feature extraction, noise reduction, noise addition, colorization were used to accentuate distinct features. Different neural network architectures, including

Table 5. The results obtained by the participating teams to the first sub-task proposed by ImageCLEFmedical GANs – Identify training data fingerprints.

Rank	Team	ID #	F1-score	Rank	Team	ID #	F1-score
#1	Inoue Koki	892	0.666	#28	csmorgan	884	0.5
#2	Inoue Koki	896	0.663	#29	AI Multimedia Lab	901	0.499
#3	Inoue Koki	891	0.663	#30	Biomedical Imaging Goa	874	0.499
#4	Inoue Koki	894	0.66	#31	Biomedical Imaging Goa	873	0.497
#5	Inoue Koki	895	0.638	#32	csmorgan	883	0.496
#6	Inoue Koki	890	0.631	#33	csmorgan	886	0.492
#7	AI Multimedia Lab	909	0.627	#34	KDE-med-lab	854	0.488
#8	Inoue Koki	893	0.626	#35	csmorgan	879	0.483
#9	SDVAHCS/UCSD	848	0.624	#36	csmorgan	878	0.47
#10	SDVAHCS/UCSD	849	0.606	#37	Shitongcao	833	0.462
#11	Robot	844	0.603	#38	KDE-med-lab	857	0.46
#12	Shitongcao	834	0.598	#39	KDE-med-lab	853	0.455
#13	Shitongcao	836	0.598	#40	KDElab	897	0.454
#14	AI Multimedia Lab	905	0.538	#41	Shitongcao	835	0.451
#15	Biomedical Imaging Goa	898	0.531	#42	Shitongcao	839	0.448
#16	Shitongcao	838	0.529	#43	KDE-med-lab	856	0.443
#17	AI Multimedia Lab	906	0.527	#44	Biomedical Imaging Goa	876	0.43
#18	Robot	841	0.524	#45	Robot	845	0.429
#19	AI Multimedia Lab	903	0.515	#46	Biomedical Imaging Goa	877	0.385
#20	Biomedical Imaging Goa	875	0.515	#47	Robot	846	0.35
#21	SDVAHCS/UCSD	850	0.511	#48	Robot	842	0.314
#22	KDE-med-lab	852	0.51	#49	Robot	843	0.312
#23	AI Multimedia Lab	904	0.51	#50	Robot	847	0.312
#24	Robot	840	0.503	#51	Shitongcao	837	0.255
#25	AI Multimedia Lab	902	0.502	#52	AI Multimedia Lab	908	0.2358
#26	SDVAHCS/UCSD	851	0.501	#53	Shitongcao	832	0.2
#27	csmorgan	881	0.5	#54	KDE-med-lab	855	0.019

ResNet, MobileNet and autoencoders were used for feature extraction and classification. The task was evaluated as a binary-class classification problem and the evaluation was carried out by measuring the F1-score, the official evaluation metric of this year's edition. The results are presented in Table 5.

For the second sub-task, "Detect generative models fingerprints", most teams used pre-trained deep learning models such as ResNet, DenseNet, EfficientNet, MobileNetV2, VGG, and Inception for feature extraction. These models were chosen for their proven efficacy in capturing complex patterns and hierarchical features in images. A variety of clustering algorithms were employed across the methods. K-means was the most commonly used clustering algorithm, but other techniques like hierarchical clustering, Gaussian Mixture Models (GMM), and t-SNE were also applied to group the extracted features based on their similarities. Many approaches involved combining multiple models or techniques to enhance robustness. Adjusted Rand Index (ARI) was the official evaluation metric of

Table 6. The results obtained by the participating teams to the second sub-task proposed by ImageCLEFmedical GANs – Detect generative models' fingerprints

Rank	Team	ID #	ARI	Rank	Team	ID #	ARI
#1	SDVAHCS/UCSD	545	1	#24	Csmorgan	451	0.267530
#2	AI Multimedia Lab	330	0.997085	#25	Csmorgan	458	0.232390
#3	AI Multimedia Lab	327	0.996517	#26	KDE-med-lab	237	0.226339
#4	AI Multimedia Lab	326	0.934709	#27	Csmorgan	456	0.178545
#5	AI Multimedia Lab	331	0.900844	#28	KDE-med-lab	248	0.166582
#6	Csmorgan	447	0.9000159	#29	KDE-med-lab	257	0.123426
#7	SDVAHCS/UCSD	550	0.885478	#30	KDE-med-lab	271	0.091818
#8	SDVAHCS/UCSD	590	0.877797	#31	KDE-med-lab	258	0.060058
#9	SDVAHCS/UCSD	548	0.851990	#32	KDE-med-lab	254	0.045286
#10	SDVAHCS/UCSD	549	0.851362	#33	KDE-med-lab	270	0.038242
#11	Csmorgan	446	0.813749	#34	KDE-med-lab	259	0.014388
#12	AI Multimedia Lab	334	0.722857	#35	KDE-med-lab	480	0.013856
#13	AI Multimedia Lab	333	0.654021	#36	SDVAHCS/UCSD	546	0.003375
#14	AI Multimedia Lab	335	0.645386	#37	Csmorgan	454	0.001776
#15	Biomedical Imaging Goa	307	0.638117	#38	Csmorgan	453	0.001313
#16	SDVAHCS/UCSD	547	0.577203	#39	KDE-med-lab	236	0.000816
#17	SDVAHCS/UCSD	225	0.577203	#40	GAN-Amis	516	0.000079
#18	AI Multimedia Lab	332	0.552682	#41	Biomedical Imaging Goa	323	0.000046
#19	AI Multimedia Lab	329	0.5037	#42	GAN-Amis	518	-0.000010
#20	Biomedical Imaging Goa	321	0.434414	#43	GAN-Amis	520	-0.000546
#21	Csmorgan	452	0.365604	#44	GAN-Amis	277	-0.000615
#22	AI Multimedia Lab	328	0.329388	#45	GAN-Amis	513	-0.000993
#23	Biomedical Imaging Goa	324	0.272975	#46	GAN-Amis	517	-0.002019

the competition and the results are presented in Table 6 More detailed results, including methods presentation and other performance measures, are presented in the overview article [4].

5.5 Lessons Learned and Next Steps

The second edition of the ImageCLEFmedical GANs task introduced two subtasks for participants: a prediction-based task utilizing both real and generated images and a clustering task using only generated images. This task provided insights into the complexities of working with synthetic medical images. Participants employed a variety of methods, including advanced image preprocessing techniques, deep learning models, and clustering algorithms for the two proposed sub-tasks.

Future editions of the task will expand the scope by incorporating a wider variety of data and generation methods to better reflect real-world applications and address existing limitations. Furthermore, new tasks will be introduced to explore different aspects of privacy and security in synthetic medical data and alternative evaluation metrics will be investigated to ensure a more comprehensive assessment of the methodologies employed.

6 The MEDVQA-GI Task

The second iteration of the MedVQA-GI challenge introduces a new goal that focuses on the use of generative models of text to image in medical diagnosis. This combines natural language processing and image generation to potentially improve diagnostic processes in healthcare by providing more comprehensive datasets that can be used for machine learning training. In contrast to last year's focus on a visual question answering task that required retrieving images or masks from user questions, this year's task aims to use generative models to create synthetic medical images from textual inputs. Participants are tasked generating the synthetic images using existing generative models developed using a dataset derived from last years MedVQA-GI challenge.

6.1 Task Setup

This year, the competition is divided into two subtasks: Image Synthesis (IS) and Optimal Prompt Generation (OPG). Participants are welcome to submit entries for one or both tasks, with no restrictions on the number of submissions.

The IS task challenged participants to use text-to-image generative models to create a dataset of medical images from textual descriptions. The objective is to produce accurate visual representations of various medical conditions described in text. For example, given the description "An early-stage colorectal polyp", participants are expected to generate an image that precisely reflects the text description.

The OPG task asked participants to build prompts that guide the generation of images meeting specific medical imaging requirements. This task tests the ability to develop prompts that result in images accurately matching predefined categories, emphasizing the model's capability to produce precise and clinically relevant images. For more comprehensive details on the tasks, datasets, and evaluation metrics, please see the task overview paper [21].

6.2 Dataset

The dataset used for this year's challenge is based on data developed for last years challenge, which is based on the HyperKvasir dataset [13] and the Kvasir-Instrument dataset [26] datasets. Participants were provided with a development dataset consisting of $2,000$ image and text pairs, and a list of $5,000$ prompts to generate their results. The development data was organized with a directory containing the images and CSV files containing the prompts and connection to the image filenames. For testing, we provided a list of prompts that participants used to generate their synthetic images.

6.3 Results

Overall, we had a total of six runs submitted to Task 1 and none to Task 2, where each team submitted three runs and the results are shown in Table 7.

Team MMCP [16] employed two distinct methods for image generation: they fine-tuned existing Kandinsky models and developed a Medical Synthesis with Diffusion Model (MSDM), with the latter showing superior results. Team 2 [31] explored three different approaches in their work. Initially, they used a CLIP model to retrieve images closely related to the input prompts rather than generating new ones. Next, they used a fine-tuned stable diffusion model for creating synthetic images. Lastly, they implemented a Low-Rank Adaptation of Large Language Models (LoRA), modifying a stable diffusion model to produce high-quality images that closely match the input specifications. Overall, the best submission goes to Team MMCP [16], who achieved best results on the quantitative metrics and also visually best results.

6.4 Lessons Learned and Next Steps

Overall, we observed a reduction in participation compared to last year. There may be several reasons for this, like the complexity of tasks, change of direction from last year, or a lack of foundational resources among the participants. Addressing these barriers could involve "getting started" scripts and potentially simplifying the challenge structure to attract a broader range of participants.

Table 7. Results for Task 1. Each submission is evaluate dusing the FID and the Inception Score (IS). The FID scores is calculated against the MedVQA testing datasert (Single), GastroVision (Multi), and a combination of the two (Both). The IS score is calculated on a 10-way split of the synthetic images, where we display the mean (avg), standard deviation (sd), and median (med).

Team	Submission	FID (Single)	FID (Multi)	FID (Both)	IS (avg)	IS (std)	IS (med)
MMCP	1	0.125	0.121	0.119	1.773	0.023	1.775
	2	0.120	0.117	0.115	1.791	0.028	1.792
	3	0.086	0.064	0.066	1.624	0.031	1.633
team2	1	0.114	0.128	0.124	1.568	0.025	1.560
	2	0.099	0.064	0.067	2.327	0.065	2.339
	3	0.110	0.073	0.076	2.362	0.050	2.359

7 ToPicto

Several diseases (e.g., Rett syndrome, Cerebral Palsy, Parkinson's Disease) lead to language impairment, which significantly interferes, as a consequence, with the development of language skills (speaking, listening, reading, and writing). Language production and comprehension are impaired. For these specific cases, Augmentative and Alternative Communication (AAC) can be implemented with the use of pictograms [39]. Pictograms, in AAC, refer to an image linked to a concept that can be a single word, a named entity, or a multi-word expression among others. Using pictograms as a communication aid has been proven effective in visualizing syntax, manipulating words, and facilitating language access [15]. Moreover, the use of AAC has a positive social impact for people with language impairment. The French Red Cross has identified a reduction in stress, an improvement in autonomy and health, and greater serenity and enjoyment in daily life. The main objective of this task is to provide a translation in pictogram terms (each linked to a specific pictogram image from the ARASAAC bank[9] from a natural language (speech or text) understandable by the users with language impairments. The translation has to follow a specific structure and should convey the meaning of the input.

7.1 Task Setup

The first edition of the ToPicto task consisted of two subtasks: *Text-to-Picto* and *Speech-to-Picto*. Participants could choose to work on both tasks or just one of them without any obligation to achieve specific results.In the Text-to-Picto subtask, participants were asked to translate a text input into a pictogram sequence. The subtask involved implementing translation techniques and models to generate a specific pictogram sequence. The second subtask, Speech-to-Picto is the continuation of Text-to-Picto, but focuses on the speech modality. Participants had to generate a pictogram sequence from a speech input. The objective was to adapt current spoken language translation systems, such has in [11] to the pictogram generation.

7.2 Dataset

The dataset consisted of oral transcriptions (for the Text-to-Picto subtask) and audio utterances (for the Speech-to-Picto subtask) translated into sequences of pictogram terms built from the TCOF corpus [2]. The TCOF corpus contains interactions between adults, adults and children, and children themselves, covering a wide range of topics such as debates, everyday situations, and medical consultations. This type of text is representative of the interactions we observe between caregivers (families, medical staff) and individuals who rely on pictograms due to language impairments.

For each utterance, we applied the method of [28] to extract the pictogram sequence. This sequence was carefully developed and evaluated by experts of the pictographic language. The audio files were a maximum of 30 s length with a sampling rate of 16 kHz. For the challenge, the dataset was split into three sets, training, validation and test with a 90/5/5 distribution respectively. General statistics about the dataset are presented in Table 8. The resulting data were provided in a JSON format to the participants with the following information: (i)*id*: the unique identifier of the utterance; (ii)*src*: the input sequence (either text or speech); (iii) *tgt*: the target sequence of pictogram terms; (iv) *pictos*: a list of pictogram identifier linked to each pictogram terms.

[9] https://arasaac.org/.

The *pictos* tag was provided for reference to give an idea of the input with the sequence of pictogram images. Each pictogram image could be obtained from the ARASAAC website from the provided identifier. The dataset will be released shortly after the end of the challenge.

Table 8. General statistics of the ToPicto dataset.

	train	valid	test
# utterances	24,270	1,348	1,350

7.3 Participating Groups and Submitted Runs

A total of 16 teams participated in the ToPicto challenge, with most registering for both tasks. Four teams completed the Text-to-Picto task. Unfortunately, no submissions were received for the Speech-to-Picto subtask. Every team provided their working notes, resulting in a 100% adherence rate.

7.4 Results

In the following section, we only discuss the submission from the Text-to-Picto subtask. The participants employed several models that are based on the same architecture, Transformer [47]. Two teams made use of multilingual pre-trained models, T5 [37] and Helsinki-NLP/opus-mt-ROMANCE-en[10]. Other models, monolingual, were also applied, specifically on the French language with CamemBERT [30] and on the English language with GPT-2 model [36]. A final work implemented an encoder-decoder architecture with LSTM layers. The evaluation was based on metrics commonly used in the translation community. The evaluation process involved comparing the reference pictogram terms sequence with the hypothesis given by the model. Three metrics were computed: BLEU score [33], METEOR [6] and the Picto-term Error Rate (PictoER), which is based on the Word Error Rate metric [49]. The results are presented in Table 9.

Table 9. The results obtained by the participating teams to the Text-to-Picto sub-task of ToPicto.

Rank	Team	Run	BLEU	METEOR	PictoER
#1	TechTitans	3	74.36	87.08	13.90
#2	TechTitans	2	67.85	83.69	17.57
#3	TechTitans	3	66.56	82.89	18.43
#4	InnoVate	2	68.96	83.54	18.51
#5	SSN-MLRG	1	3.41	14.35	141.90
#6	SSN-MLRG	2	3.41	14.35	141.90
#7	InnoVate	2	3.93	25.56	170.80

[10] https://huggingface.co/Helsinki-NLP/opus-mt-ROMANCE-en.

7.5 Lessons Learned and Next Steps

The first edition of the task introduced two subtasks: generating a coherent sequence of pictogram terms from either a text utterance (Text-to-Picto) or a speech utterance (Speech-to-Picto). This challenge, previously receiving limited attention, was presented to the community for the first time. Participants employed a variety of methods, ranging from multilingual to monolingual pre-trained models, and encoder-decoder architectures, yielding interesting outcomes in translation. However, the Speech-to-Picto subtask did not result in any submissions, likely due to the challenges associated with starting from a speech modality.

Future editions of the task might explore different language sets and various domains, such as the medical field. Additionally, an important aspect of providing comprehensible translations is simplifying the text input beforehand, which could serve as a new subtask in the ToPicto challenge. Finally, the dynamic construction of pictograms using generative models could also be explored.

8 Conclusion

This paper presents an overview and the outcomes of ImageCLEF 2024 benchmarking campaign. Three main tasks were organised, addressing challenges in the medical domain (caption analysis, visual question answering, medical dialogue summarisation, GANs for medical image generation), natural language translation (generating pictogram from speech and text), and image retrieval/-generation for arguments.

Similar to the previous year, the vast majority of solutions provided by the participants were based on machine learning and deep learning techniques. In ImageCLEFmedical – Caption, multi-label classification was common for concept detection, with some teams integrating image retrieval. Encoder-decoder frameworks with transformers and LSTMs were used for caption prediction. In ImageCLEFmedical – MEDIQA-MAGIC, participants used classic algorithms like SVM, KNN, and Random Forest, along with TF-IDF and lemmatization. Pre-trained models like GPT3.5, clinical-BERT, and clinical T5, including their LoRA adaptations, were also utilized. For ImageCLEFmedical GAN, methods included advanced preprocessing, deep learning models, binarization, histogram equalization, and feature extraction. Majority voting and agglomerative clustering improved results. For the second sub-task, pre-trained CNNs were used for feature extraction, with clustering algorithms like k-means, hierarchical clustering, GMM, and t-SNE. For the ImageCLEFmedical-MedvQA, the participants employed transformer-based pre-trained models. In the first edition of the ToPicto task, methods for Text-to-Picto included multilingual and monolingual pre-trained models, and encoder-decoder architectures, achieving interesting translation outcomes. ImageCLEF 2024 offered participants and the community a wide range of tasks and methodologies to delve into, highlighting an exciting fusion of approaches.

Future editions of the ImageCLEF tasks hold exciting potential for growth and innovation. They may broaden domains, including tasks to attract more people, and try new methods like generative models for the GANs task. To overcome barriers in participation, like complicated taks, offering resources may be necessary. Additionally, refining evaluation metrics and exploring alternative approaches are crucial for advancing understanding across disciplines. These actions aim to drive progress and foster collaboration in diverse areas of research.

Acknowledgements. The lab is supported under the H2020 AI4Media "A European Excellence Centre for Media, Society and Democracy" project, contract #951911, as well as the ImageCLEFmedical GANs tasks. The work of Louise Bloch, Raphael Brüngel and Benjamin Bracke was partially funded by a PhD grant from the University of Applied Sciences and Arts Dortmund (FH Dortmund), Germany. The work of Ahmad Idrissi-Yaghir, Tabea M. G. Pakull, Hendrik Damm and Henning Schäfer was funded by a PhD grant from the DFG Research Training Group 2535 Knowledge- and data-based personalisation of medicine at the point of care (WisPerMed). The ToPicto task was funded by the Agence Nationale de la Recherche (ANR) through the project PROPICTO (ANR-20-CE93-0005).

References

1. Gemini models (2024). https://ai.google.dev/gemini-api/docs/models/gemini. Accessed 24 Apr 2024
2. André, V., Canut, E.: Mise à disposition de corpus oraux interactifs: le projet tcof (traitement de corpus oraux en français). Pratiques. Linguistique, littérature, didactique (147-148), 35–51 (2010)
3. Andrei, A., Radzhabov, A., Coman, I., Kovalev, V., Ionescu, B., Müller, H.: Overview of ImageCLEFmedical GANs 2023 task – identifying training data "fingerprints" in synthetic biomedical images generated by GANs for medical image security. In: CLEF2023 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 18–21 September 2023
4. Andrei, A., et al.: Overview of 2024 ImageCLEFmedical GANs task – investigating generative models' impact on biomedical synthetic images. In: CLEF2024 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 9–12 September 2024
5. Banerjee, S., Lavie, A.: Meteor: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72. Association for Computational Linguistics, Ann Arbor, Michigan, June 2005. https://aclanthology.org/W05-0909
6. Banerjee, S., Lavie, A.: Meteor: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72 (2005)

7. Ben Abacha, A., Mrabet, Y., Zhang, Y., Shivade, C., Langlotz, C.P., Demner-Fushman, D.: Overview of the MEDIQA 2021 shared task on summarization in the medical domain. In: Proceedings of the 20th Workshop on Biomedical Language Processing, BioNLP@NAACL-HLT 2021, Online, 11 June 2021, pp. 74–85. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.bionlp-1.8
8. Ben Abacha, A., Wai Yim, W., Adams, G., Snider, N., Yetisgen, M.: Overview of the MEDIQA-chat 2023 shared tasks on the summarization and generation of doctor-patient conversations. In: ACL-ClinicalNLP 2023 (2023)
9. Ben Abacha, A., et al.: Overview of the MEDIQA-M3G 2024 shared tasks on multilingual multimodal medical answer generation. In: NAACL-ClinicalNLP 2024 (2024)
10. Ben Abacha, A., Yim, W., Michalopoulos, G., Lin, T.: An investigation of evaluation methods in automatic medical note generation. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, pp. 2575–2588. Association for Computational Linguistics, July 2023. https://doi.org/10.18653/v1/2023.findings-acl.161. https://aclanthology.org/2023.findings-acl.161
11. Bérard, A., Besacier, L., Kocabiyikoglu, A.C., Pietquin, O.: End-to-end automatic speech translation of audiobooks. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6224–6228. IEEE (2018)
12. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. **32**(Database-Issue), 267–270 (2004). https://doi.org/10.1093/nar/gkh061
13. Borgli, H., et al.: Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Sci. Data **7**(1) (2020). https://doi.org/10.1038/s41597-020-00622-y
14. Carmo, D., Rittner, L., Lotufo, R.: VisualT5: multitasking caption and concept prediction with pre-trained ViT, T5 and customized spatial attention in radiological images. In: CLEF2024 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 9–12 September 2024
15. Cataix-Nègre, E.: Communiquer autrement: Accompagner les personnes avec des troubles de la parole ou du langage. De Boeck Superieur (2017)
16. Chaychuk, M.: MMCP team at ImageCLEFmed 2024 task on image synthesis: diffusion models for text-to-image generation of colonoscopy images. In: CLEF2024 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, September 2024
17. Galley, M., et al.: deltaBLEU: a discriminative metric for generation tasks with intrinsically diverse targets. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 445–450. Association for Computational Linguistics, Beijing, China, July 2015
18. García Seco de Herrera, A., Schaer, R., Bromuri, S., Müller, H.: Overview of the ImageCLEF 2016 medical task. In: Working Notes of CLEF 2016 (Cross Language Evaluation Forum), September 2016
19. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: a reference-free evaluation metric for image captioning. In: Moens, M., Huang, X., Specia, L., Yih, S.W. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana,

Dominican Republic, 7–11 November 2021, pp. 7514–7528. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.emnlp-main.595. https://doi.org/10.18653/v1/2021.emnlp-main.595
20. Hicks, S.A., Storås, A., Halvorsen, P., de Lange, T., Riegler, M.A., Thambawita, V.: Overview of ImageCLEFmedical 2023 - medical visual question answering for gastrointestinal tract. In: CLEF2023 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, September 2023
21. Hicks, S.A., Storås, A., Halvorsen, P., Riegler, M.A., Thambawita, V.: Overview of ImageCLEFmedical 2024 - medical visual question answering for gastrointestinal tract. In: CLEF2024 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, September 2024
22. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
23. Ionescu, B., et al.: Advancing multimedia retrieval in medical, social media and content recommendation applications with ImageCLEF 2024. In: Goharian, N., et al. (eds.) ECIR 2024. LNCS, pp. 44–52. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-56072-9_6
24. Ionescu, B., et al.: Overview of ImageCLEF 2023: multimedia retrieval in medical, socialmedia and recommender systems applications. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023). Springer Lecture Notes in Computer Science LNCS, Thessaloniki, Greece, 18–21 September 2023
25. Ionescu, B., et al.: ImageCLEF 2019: multimedia retrieval in medicine, lifelogging, security and nature. In: Crestani, F., et al. (eds.) CLEF 2019. LNCS, vol. 11696, pp. 358–386. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28577-7_28
26. Jha, D., et al.: Kvasir-instrument: diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. In: Lokoč, J., et al. (eds.) MMM 2021. LNCS, vol. 12573, pp. 218–229. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67835-7_19
27. Kiesel, J., et al.: Overview of touché 2024: argumentation systems. In: Goeuriot, L., et al. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024). Lecture Notes in Computer Science, Springer, Berlin (2024)
28. Macaire, C., et al.: A multimodal French corpus of aligned speech, text, and pictogram sequences for speech-to-pictogram machine translation. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 839–849. ELRA and ICCL, Torino, Italia, May 2024. https://aclanthology.org/2024.lrec-main.76
29. Macaire, C., Esperança-Rodier, E., Lecouteux, B., Schwab, D.: Overview of ImageCLEFToPicto 2024 – investigating the translation of natural language into pictograms. In: CLEF2024 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 9–12 September 2024
30. Martin, L., et al.: CamemBERT: a tasty French language model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7203–7219. Association for Computational Linguistics, Online, July 2020. https://www.aclweb.org/anthology/2020.acl-main.645
31. Oluwafemi Ojonugwa, E.P., Rahman, M., Khalifa, F.: Advancing AI-powered medical image synthesis: insights from MEDVQA-GI challenge using clip, fine-tuned stable diffusion, and dream-booth + LoRA. In: CLEF2024 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, September 2024

32. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 311–318. Association for Computational Linguistics, July 2002. https://doi.org/10.3115/1073083.1073135. https://aclanthology.org/P02-1040
33. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)
34. Pelka, O., Friedrich, C.M., García Seco de Herrera, A., Müller, H.: Overview of the ImageCLEFmed 2020 concept prediction task: medical image understanding. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 22–25 September 2020
35. Popescu, A., Deshayes-Chossart, J., Schindler, H., Ionescu, B.: Overview of the ImageCLEF 2022 aware task. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 13th International Conference of the CLEF Association (CLEF 2022), Bologna, Italy. LNCS. Springer, Cham, 5–8 September 2022
36. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)
37. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(140), 1–67 (2020). http://jmlr.org/papers/v21/20-074.html
38. Roberts, R.J.: PubMed central: the GenBank of the published literature. Proc. Natl. Acad. Sci. U.S.A. **98**(2), 381–382 (2001). https://doi.org/10.1073/pnas.98.2.381
39. Romski, M., Sevcik, R.A.: Augmentative communication and early intervention: myths and realities. Infants Young Child. **18**(3), 174–185 (2005)
40. Rückert, J., et al.: Overview of ImageCLEFmedical 2024 – caption prediction and concept detection. In: CLEF2024 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 9–12 September 2024
41. Rückert, J., et al.: Overview of ImageCLEFmedical 2023 – caption prediction and concept detection. In: CLEF2023 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 18–21 September 2023
42. Rückert, J., et al.: ROCOv2: radiology objects in context version 2, an updated multimodal image dataset. Sci. Data (2024). https://doi.org/10.1038/s41597-024-03496-6. https://arxiv.org/abs/2405.10004v1
43. Sellam, T., Das, D., Parikh, A.P.: BLEURT: learning robust metrics for text generation. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, 5–10 July 2020, pp. 7881–7892. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.704. https://doi.org/10.18653/v1/2020.acl-main.704
44. Ştefan, L.D., Constantin, M.G., Dogariu, M., Ionescu, B.: Overview of ImageCLEFfusion 2023 task - testing ensembling methods in diverse scenarios. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 18–21 September 2023
45. Tsikrika, T., de Herrera, A.G.S., Müller, H.: Assessing the scholarly impact of ImageCLEF. In: Forner, P., Gonzalo, J., Kekäläinen, J., Lalmas, M., de Rijke, M. (eds.) CLEF 2011. LNCS, vol. 6941, pp. 95–106. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23708-9_12

46. Tsikrika, T., Larsen, B., Müller, H., Endrullis, S., Rahm, E.: The scholarly impact of CLEF (2000–2009). In: Forner, P., Müller, H., Paredes, R., Rosso, P., Stein, B. (eds.) CLEF 2013. LNCS, vol. 8138, pp. 1–12. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40802-1_1
47. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017). https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
48. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: consensus-based image description evaluation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015, pp. 4566–4575. IEEE Computer Society (2015). https://doi.org/10.1109/CVPR.2015.7299087
49. Woodard, J., Nelson, J.: An information theoretic measure of speech recognition performance. In: Workshop on Standardisation for Speech I/O Technology, Naval Air Development Center, Warminster, PA (1982)
50. Yim, W., Ben Abacha, A., Fu, Y., Sun, Z., Yetisgen, M., Xia, F.: Overview of the MEDIQA-magic task at ImageCLEF 2024: multimodal and generative telemedicine in dermatology. In: CLEF 2024 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 9–12 September 2024
51. Yim, W., Ben Abacha, A., Snider, N., Adams, G., Yetisgen, M.: Overview of the MEDIQA-sum task at ImageCLEF 2023: summarization and classification of doctor-patient conversations. In: CLEF 2023 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 18–21 September 2023
52. Yim, W., Fu, Y., Sun, Z., Ben Abacha, A., Yetisgen, M., Xia, F.: DermaVQA: a multilingual visual question answering dataset for dermatology. CoRR (2024)
53. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTscore: evaluating text generation with BERT. arXiv abs/1904.09675 (2019)
54. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTscore: evaluating text generation with BERT. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 26–30 April 2020. OpenReview.net (2020). https://openreview.net/forum?id=SkeHuCVFDr

Overview of the CLEF 2024 JOKER Track

Automatic Humour Analysis

Liana Ermakova[1(✉)], Anne-Gwenn Bosser[2], Tristan Miller[3,4], Victor Manuel Palma Preciado[1,5], Grigori Sidorov[5], and Adam Jatowt[6]

[1] Université de Bretagne Occidentale, HCTI, Brest, France
liana.ermakova@univ-brest.fr
[2] École Nationale d'Ingénieurs de Brest, Lab-STICC CNRS UMR 6285, Plouzané, France
[3] Department of Computer Science, University of Manitoba, Winnipeg, Canada
[4] Austrian Research Institute for Artificial Intelligence (OFAI), Vienna, Austria
[5] Centro de Investigación en Computación (CIC), Instituto Politécnico Nacional (IPN), Mexico City, Mexico
[6] University of Innsbruck, Innsbruck, Austria

Abstract. The JOKER Lab series at the Conference and Labs of the Evaluation Forum (CLEF) was established in 2022 to promote collaborative, interdisciplinary research on the automated processing of wordplay and verbal humour. This paper provides an overview of the setup and results of the Lab's 2024 edition. We describe the data and evaluation metrics used for the Lab's three shared tasks (on humour-aware information retrieval, humour classification according to genre and technique, and translation of puns from English to French) and introduce and compare the systems that participated in each task, with particular attention to their approaches and performance.

Keywords: Wordplay · Puns · Humour · Humour retrieval · Humour translation · Humour classification · Information retrieval

1 Introduction

Despite great strides made in recent years by large language models (LLMs), the development of fully automatic systems for the interpretation, analysis, generation and translation of humour remains a challenge. Now in its third year, the JOKER series of evaluation labs[1] at the Conference and Labs of the Evaluation Forum (CLEF) aims to make headway on this problem, in large part by bringing together researchers from AI and the social sciences, particularly linguistics. In each edition of JOKER, we construct and publish reusable, quality-controlled data sets for use as training and test data in various humour processing tasks.

The first JOKER lab, held at CLEF 2022, featured shared tasks on the categorisation and translation of wordplay, puns, and humorous neologisms, in English and French [16,17]. The next iteration of JOKER, at CLEF 2023, had

[1] https://www.joker-project.com/.

L. Goeuriot et al. (Eds.): CLEF 2024, LNCS 14959, pp. 165–182, 2024.
https://doi.org/10.1007/978-3-031-71908-0_8

Table 1. Statistics on the runs submitted to the CLEF JOKER 2024 track per task

Team	Task 1	Task 2	Task 3	Total
jokester [2]	1	1	1	3
LIS [21]	1			1
Arampatzis	10	8	6	24
Frane	1	1	1	3
AB&DPV [35]	1	7	1	9
Dajana&Kathy	1	1	1	3
Petra&Regina [8]	1	1	1	3
Tomislav&Rowan [24]	1	3	2	6
UAms [5]	8	1	2	11
RubyAiYoungTeam	1	1		2
ORPAILLEUR [9]		9		9
NaiveNeuron [22]		3		3
HumourInsights [32]		1		1
CYUT [36]		3		3
CodeRangers [25]		2		2
VayamSolveKurmaha [3]		2		2
DadJokers [30]		3		3
NLPalma [26]		3		3
PunDerstand [7]		4		4
Olga			3	3
Farhan [6]			2	2
UBO			3	3
Total	26	54	23	103

tasks on detection, location, and interpretation of puns in English, French, and Spanish [11,12], as well as on machine translation of wordplay from English into French and English into Spanish [13–15]. This year's JOKER lab featured a mix of familiar and new tasks:

Task 1 Humour-aware information retrieval
Task 2 Humour classification according to genre and technique
Task 3 Translation of puns from English to French

This paper provides an overview of the entire CLEF 2024 JOKER track. For more detailed discussions of each of the individual tasks, we refer to the detailed CLEF 2024 JOKER Task Overview papers in the CEUR proceedings. Specifically, Task 1 on retrieving [18], Task 2 on classifying [27], and Task 3 on translating [19] humourous texts.

In all, 53 teams registered for our JOKER track at CLEF 2024. Of these, 22 teams participated in the tasks, submitting a total of 103 runs for the numbered shared tasks. Statistics for these runs are presented in Table 1.

The rest of this overview is structured as follows. In the next three sections, we discuss each of the JOKER track's tasks: Task 1 on retrieving humourous texts in Sect. 2, Task 2 on classifying humourous texts in Sect. 3, and Task 3 on translating humourous texts in Sect. 4. We end with discussion and conclusions in Sect. 5.

2 Task 1: Humour-Aware Information Retrieval

This section discusses the JOKER Track's Task 1 on retrieving humourous texts.

2.1 Description

As we have shown previously, state-of-the-art AI models are wordplay- and humour-agnostic [10,11,14]. To foster research in humour-aware information retrieval, in JOKER 2024 we have introduced a new task that aims at retrieving short humorous texts from a document collection. The intended use case is to search for a joke on a specific topic. This can be useful for humour researchers in the humanities, for second-language learners as a learning aid, for professional comedians as a writing aid, and for translators who might need to adapt certain jokes to other cultures.

For this task, the aim is to retrieve short humorous texts from a document collection based on a given query. The retrieved texts should be fulfill the dual criteria of being relevant to the query and being instances of wordplay. The typical use case would be searching for a joke on a specific topic – e.g., a query of `math` means that the goal is to find math jokes, while the query `Tom` means that the goal is to find jokes about Tom.

2.2 Data

The data for this task is an extension of that used for JOKER 2023's tasks on wordplay detection in English [11], which is annotated according to whether the texts are humorous. We also included texts from the translation Task 3 of JOKER 2023 [13] and new wordplay instances used in Task 3 this year.

We extended this data by introducing retrieved text passages from non-humour sources as well as generated data on topics relevant to the queries. In particular, the non-humorous data was drawn from:

- sentences from Wikipedia extracts returned for queries using the Wikipedia Python package,[2] and
- definitions of the query terms generated by Meta's Llama 2 with 7B parameters [33].

We considered these texts to be topically relevant to the corresponding queries.

This methodology was adopted in part to avoid data artefacts related to unbalanced topics, differences in vocabulary of humourous and non-humourous

[2] https://pypi.org/project/wikipedia/.

texts, and differences between machine- and human-produced documents. The application of these data augmentation techniques was also intended prevent participants from simply mining previously released data to or performing web scraping, thereby reducing the potential for reliance on external sources or overfitting to existing corpora.

The total number of documents in the corpus is 61,268, with 4,492 humorous texts (3,507 from the JOKER 2023 wordplay detection corpus and 985 new wordplay instances) and 56,776 non-humorous ones (4,954 negative examples from the JOKER 2023 wordplay detection corpus, 12,523 texts generated by Llama 2, and 39,299 sentences from Wikipedia extracts). The texts were typically one or two sentences in length, and were provided as JSON files.

To construct queries, we used the locations of wordplay from JOKER 2023's Task 2 [12]. We provided 12 queries with judgment (qrels) to train/validate systems. Forty-five queries were used for evaluation. Note that we included all training-set queries in the test input file, but they are excluded from the resulting scores. For 57 queries (test and training), 11,831 documents were considered topically relevant. We considered the documents topically relevant to the query if the text contained the term from this query or its synonyms or hyperonyms coming from the pun interpretation annotation of JOKER 2023 Task 2 [12]. Of topically relevant documents, 1,730 are considered to be humorous.

2.3 Evaluation

For evaluation, we used standard information retrieval metrics implemented in the `pyterrier` platform [23,34]:

map mean average precision – i.e., the mean of the average precision scores for each query

ndcg normalised discounted cumulative gain, the gain of each document based on its relevance, discounted logarithmically by its position in the ranking normalised over the ideal ranking

P1, P5, P10 precision – i.e. the ability of a system to present only relevant items, at different levels

R5, R10, R100, R1000 recall – i.e., the ability of a system to present all relevant items, at different levels

bpref binary preference, a sum-based metric showing how many relevant documents are ranked before irrelevant documents

MRR mean reciprocal rank, the average of the multiplicative inverse of the ranks of the first correct answer of results for a sample of queries

2.4 Participants' Approaches

The **jokester** team [2] submitted a single run. The authors provide a simple baseline approach that uses TF–IDF for feature weighting coupled with a Logistic Regression classifier.

The **Frane** team submitted one run. The team approached humour-aware information retrieval by using fine-tuned BERT models coupled with traditional retrieval models like BM25. The texts were ranked based on their relevance as computed by BM25 and humourousness of the content assessed using BERT.

The **UAms** team [5] submitted a total of eight runs. The first two runs were baselines focussing on regular information retrieval effectiveness using either BM25 or BM25+RM3 with default settings. The next two runs were neural cross-encoder rerankings of the former runs based on zero-shot application of an MSMARCO-trained ranker, reranking the top 100 of either the BM25 or the BM25+RM3 baseline run. The remaining four runs aimed to take the pun detection of the results into account. The team trained two versions of the SimpleT5 model, one with a batch size of 6 and the other with a batch size of 8, and then trained a BERT model using LoRa.

The **LIS** team [21] submitted one run. The team's method, which used a T5 transformer model, involved processing queries, expanding them with synonyms collected from WordNet, finding the best method for tokenization of queries and documents, and then choosing the optimal threshold for the similarity score, followed finally by applying a pre-trained model to filter texts with puns.

The **AB&DPV** team [35] submitted one run. The team used TF–IDF for ranking humourous text within the collection.

The **Dajana&Kathy** team submitted one run. After thorough text processing including stemming, lemmatisation, and stop word removal, the team employed TF–IDF and BM25, followed by fine-tuning using BERT.

The **Petra&Regina** team [8] submitted one run. Their approach employs logistic regression over TF–IDF vectorised documents and queries in order to find relevant outputs with iterative relevance scoring.

The **Tomislav&Rowan** team [24] submitted a single run. Their approach was based on logistic regression with TF–IDF vectorised applied to documents.

The **Arampatzis** team (No paper received) submitted 10 runs for this task. They tested a range of different models including TF–IDF, LSTM, Random Forest, XGBoost, LightGBM (Light Gradient-Boosting Machine), SVM, Decision Tree, Gaussian Naive Bayes, KNN, and a run based on applying neural nets.

Finally, the **RubyAiYoungTeam** team submitted a single run; however, they provided no description of their method.

We do not detail the 0-scored runs and the runs with problems that we could not resolve.

2.5 Results

Ten teams submitted 26 runs for Task 1. Unfortunately, the majority of runs had problems, for example partial runs for the train data only. Table 2 presents the participants' results. As noted above, we omitted the 0-scored runs and the runs with problems that we could not resolve. We observe that the runs based on pseudo-relevance feedback RM3 query expansion outperform the BM25 baselines. Cross-encoder rerankers have not shown better performance than the base-

Table 2. JOKER Task 1 results of the participants

run ID	map	ndcg	R5	R10	R100	R1000	bpref	MRR	P1	P5	P10
UAms_rm3_T5_Filter2	.12	.28	.09	.15	.36	.43	.18	.26	.13	.11	.13
UAms_rm3_BERT_Filter	.12	.27	.09	.14	.35	.42	.16	.27	.16	.11	.12
UAms_rm3_T5_Filter1	.11	.27	.09	.15	.36	.42	.16	.23	.11	.09	.11
UAms_bm25_BERT_Filter	.09	.24	.06	.12	.37	.40	.12	.19	.09	.05	.08
AB&DPV_TFIDF	.09	.24	.07	.13	.33	.37	.10	.25	.13	.12	.14
UAms_Anserini_rm3	.08	.27	.06	.08	.38	.50	.09	.20	.11	.06	.06
jokester_1_TFIDF_LogRegr	.08	.19	.09	.09	.10	.16	.21	.51	.44	.23	.14
UAms_Anserini_bm25	.08	.24	.06	.08	.37	.42	.09	.19	.11	.05	.06
UAms_bm25_CE100	.04	.17	.03	.04	.37	.37	.06	.08	.00	.04	.03
UAms_rm3_CE100	.04	.18	.03	.04	.38	.38	.06	.07	.00	.04	.03
LIS_MiniLM-T5	.02	.05	.03	.04	.05	.05	.05	.13	.04	.06	.04

line models. Filtering trained on the wordplay detection task, however, largely improved systems' results. In general, we observe that both precision and recall are extremely low. Low precision is due to the presence of the query terms in the non-humorous texts which is considered as topical relevance by the retrieval systems. Low recall is probably related to the length of the text and the fact that in many texts, both humorous and topically relevant, the query terms do not appear.

3 Task 2: Humour Classification

This section discusses the JOKER Track's Task 2 on classifying humourous text according to genre and technique.

3.1 Description

Classification of humour is an important task in dialogue systems as it can be used to provide an appropriate answer to a playful request [31]. In this task, systems were expected to automatically classify texts according to the following humour techniques:

IR: Irony relies on a gap between the literal meaning and the intended meaning, creating a humorous twist or reversal.

SC: Sarcasm involves using irony to mock, criticise, or convey contempt.

EX: Exaggeration involves magnifying or overstating something beyond its normal or realistic proportions.

AID: Incongruity/Absurdity refers (in the case of incongruity) to the unexpected or contradictory elements that are combined in a humorous way and (for absurdity) involve presenting situations, events, or ideas that are inherently illogical, irrational, or nonsensical.

SD: Self-deprecating humour involves making fun of oneself or highlighting one's own flaws, weaknesses, or embarrassing situations in a lighthearted manner.

WS: Wit/Surprise refers (in the case of wit) to clever, quick, and intelligent humour, and (for surprise) to introducing unexpected elements, twists, or punchlines that catch the audience off guard.

Thus, the humour classification of Task 2 is a classification task where the goal is to identify in a target text the particular technique used for generating humour. Runs for this task were evaluated according to standard metrics for classification.

3.2 Data

The data for this task is a mixture of existing corpora on irony and sarcasm detection [1,20] and on COVID-19 humour [4], our JOKER corpus 2023 [14] as well as jokes retrieved from public humour sites according to the predefined categories selected in a balanced manner. An example data instance is given below:

Sentence "Finally figured out the reason I look so bad in photos. It's my face".
Humour technique Self-deprecating (SD)

For Task 2, there are 1,742 sentences in the training set all labeled as either SC, EX, WS, SD, AID, IR, or WT (as discussed above). The test data consists of 6,642 unlabeled texts that contain one of the earlier described types of humor. From these texts, 722 were used for the evaluation. The details of the support (column S) for each class are given in Table 3.

3.3 Participants' Approaches

The **AB&DPV** team [35] submitted a total of seven runs, opting to use embeddings with the help of Word2Vec. To develop their results, they used the Multilayer Perceptron (MLP), Random Forest, Decision Tree, and Gaussian Naive Bayes classifiers. Their final testing accuracy ranged from 39% to 48%.

The **CodeRangers** team [25] submitted a total of two runs. The team used BERT-uncased and RoBERTa by fine-tuning them on the provided data for Task 2. Employing RoBERTa and BERT-uncased for classification, with an 80/20 setup, resulted in slightly higher accuracy with RoBERTa at 67.05% compared to BERT's 66.56% during their experiments on the training data.

The **CYUT** team [36] submitted a total of three runs. RoBERTa was first fine-tuned, using an 80/20 split of the dataset we shared to train and validate models, with a 71.63% accuracy result. However, the results on our test set are low. GPT-4 was used with zero-shot prompting and chain-of-thought prompting. Attempts to classify among all the classes at once proved too challenging for the model. Therefore hierarchical categories were created and a four-step classification method was employed (using either binary or three-way classification for a given step). After first discriminating AID and WS on one hand, from IR, SC, SD, and EX on the other, further steps allowed to classify down the grouping hierarchy. Llama 3-8b was fine-tuned on a single GPU, which was made possible by utilising four-bit quantization with QLoRa.

The **DadJokers** team [30] submitted a total of three runs. The team used BERT and a traditional machine learning model such as a Random Forest classifier. The first classification approach is done using BERT base uncased and the second attempt for classification is through using a Random Forest classifier. The authors applied `TFIDFVectorizer` and `SentenceTransformer` as preprocessing steps.

The **Dajana&Kathy** team submitted a single run. Their approach involves the use of TF–IDF and BERT embeddings with a variety of models such as SVM, Random Forest, LSTM, and Transformers. A similar approach was taken by the **Frane** team.

The **Jokester** team [2] submitted a single run. They combined several classifiers available through the Scikit library: a voting classifier weighted the results obtained by an SVC and a stack of Random Forest, Decision Tree, Gradient Boosting, and Logistic Regression. We do not report their results as all texts were predicted the SD class.

The **HumourInsights** team [32] submitted one run. Although this team reported using a variety of classical approaches for the classification task, they chose to submit only their best model. They used TF–IDF to extract features for later use in different methods. They employed boosting methods such as ADA and Gradient with mixed results, but these did not reach the accuracy obtained with KNN and Random Forest.

The **NLPalma** team [26] made 3 submissions for the classification task, using two different approaches: one with more classical classifiers and the other with a more well-known model in the BERT-like lineage.

The **PunDerstand** team [7] submitted a total of four runs. The authors employed the DeBERTa model which, after fine-tuning, gave rise to two runs, one on a raw, unprocessed dataset and one on balanced data. The latter was ensured by undersampling strategy. In another run, they used GPT-4o, the most recent large language model developed by OpenAI. Few-shot prompting was employed with one example for each class of humor. The team also provided a run with manually guided annotation.

The **Tomislav&Rowan** team [24] submitted a total of three runs. After preprocessing the text, TF–IDF was used for vectorising it. Three models were trained on the data: Logistic Regression, Naïve Bayes, and an SVM.

The **Petra&Regina** team [8] submitted a total of one run. Processed data was vectorised using TF–IDF and class labels were encoded using a linear regression algorithm.

The **UAms** team [5] submitted a single run. This team chose to use a BERT classifier trained on 90% of the train data, with taking special precautions against overfitting.

The **ORPAILLEUR** team [9] submitted a total of nine runs. The team explored the potential of advanced LLMs within a consistent methodological framework. They employed a four-bit quantised version of three LLMs: Llama2-7b1, Mistral-7b2, and Llama3-8b. The final hidden state of the last token was used as input to the feed-forward layer with the softmax function to get the class probabilities. The team also explored QLoRa adapters.

The **NaiveNeuron** team [22] submitted three runs through iterations of different LLMs, including various instances of GPT-3.5, GPT-4, and GPT-4.0 Plus RAG. They obtained the best results with GPT-4+RAG using a 70/15/15 split for testing. The experimentation of this team was not limited to GPT models; they also used fastText and Llama 3. However, they achieved more favorable results with zero-shot and few-shot classification using GPT-RAG.

The **VayamSolveKurmaha** team [3] submitted a pair of runs, primarily using LaBSE for the embeddings and BERT for the classification of this task.

The **Arampatzis** team submitted eight runs for this task. The team has experimented with the following approaches: XLNet, Multilayer Perceptron, BERT, RoBERTa, DistilBERT, DeBERTa, Electra, AlBERT.

Finally, the **RubyAiYoungTeam** team submitted a single run, without providing details of their approach.

3.4 Results

Eighteen teams submitted 54 runs for Task 2. This was the most popular JOKER task this year which might be explained by the variety of classification models. Participants used mostly LLMs and traditional classifiers although some teams experimented with fine-tuned models and different setups. Table 3 presents the results; for each run and each class, we report precision (P), recall (R), F_1, and the number of instances in this class (S – support). We also report accuracy (Acc), macro and weighted average precision, recall, and F_1 for each run. LLM-based models clearly outperform traditional classifiers. The most difficult classes were Exaggeration (EX) and Wit/Surprise (WS). The latter category is a combination of two types of humour which might be a reason for its difficulty. Further analysis is needed.

4 Task 3: Pun Translation

This section discusses the JOKER Track's Task 3 on translating puns from English to French.

4.1 Description

This year, we continued to hold the pun translation task as in JOKER 2022 [17] and 2023 [14]. The goal of this task is to translate English punning jokes into French. Translations should aim to preserve, to the extent possible, both the form and meaning of the original wordplay.

Table 3. Task 2 results of the participants: Acc (accuracy), and P (precision), R (recall), and F1 per class and macro and weighted average. S is the number of instances.

run_id	Acc	SD				WS				EX				IR				SC				AID				macro avg			weighted avg			
		P	R	F_1	S	P	R	F_1	S	P	R	F_1	S	P	R	F_1	S	P	R	F_1	S	P	R	F_1	S	P	R	F_1	P	R	F_1	S
ORPAILLEUR_mistral-7b-ens	.76	.76	.78	.77	91	.61	.51	.56	49	.64	.43	.52	106	.67	.83	.74	147	.73	.75	.74	59	.87	.88	.87	270	.71	.70	.70	.75	.76	.75	722
ORPAILLEUR_llama2-ens	.74	.77	.82	.80	91	.55	.45	.49	49	.64	.25	.36	106	.60	.78	.68	147	.65	.81	.72	59	.89	.91	.90	270	.68	.67	.66	.74	.74	.72	722
ORPAILLEUR_llama3-8b-ens	.73	.76	.85	.80	91	.54	.39	.45	49	.48	.41	.44	106	.59	.72	.65	147	.76	.66	.71	59	.90	.89	.90	270	.67	.65	.66	.72	.73	.72	722
ORPAILLEUR_mistral-7b-high	.72	.71	.75	.73	91	.53	.41	.46	49	.58	.47	.52	106	.65	.77	.70	147	.69	.69	.69	59	.86	.85	.85	270	.67	.66	.66	.72	.72	.72	722
ORPAILLEUR_llama2-high	.71	.76	.81	.78	91	.45	.37	.40	49	.51	.49	.50	106	.67	.59	.63	147	.52	.73	.61	59	.89	.89	.89	270	.63	.65	.64	.71	.71	.71	722
CodeRangers_roberta	.70	.37	.66	.47	38	.58	.87	.70	98	.85	.64	.73	44	1	.01	.03	74	.86	.65	.74	65	.83	.93	.88	190	.75	.63	.59	.78	.70	.66	509
ORPAILLEUR_llama2-low	.70	.82	.71	.76	91	.46	.43	.44	49	.63	.21	.31	106	.55	.74	.63	147	.57	.83	.68	59	.86	.89	.87	270	.65	.64	.62	.71	.70	.68	722
ORPAILLEUR_llama3-8b-high	.70	.75	.85	.79	91	.41	.24	.31	49	.45	.37	.40	106	.56	.67	.61	147	.73	.64	.68	59	.87	.90	.88	270	.63	.61	.61	.69	.70	.69	722
ORPAILLEUR_llama3-8b-low	.70	.77	.79	.78	91	.46	.37	.41	49	.49	.39	.43	106	.54	.65	.59	147	.68	.68	.68	59	.88	.88	.88	270	.64	.63	.63	.70	.70	.70	722
CYUT_llama3-FT	.69	.7	.69	.69	91	.44	.63	.52	49	.51	.40	.45	105	.63	.60	.61	146	.66	.67	.67	59	.85	.88	.86	268	.63	.64	.63	.69	.69	.69	718
PunDerstand_DeBERTa	.69	.68	.90	.77	91	.45	.51	.48	49	.43	.19	.26	106	.59	.60	.60	147	.50	.83	.62	59	.92	.86	.89	270	.59	.65	.60	.68	.69	.67	722
Arampatzis_BERT	.68	.74	.82	.78	91	.44	.31	.36	49	.50	.29	.37	106	.55	.69	.62	147	.51	.59	.55	59	.86	.87	.86	270	.60	.60	.59	.67	.68	.67	722
Arampatzis_deberta	.68	.75	.84	.79	91	.41	.55	.47	49	.48	.25	.33	106	.58	.54	.56	147	.58	.66	.62	59	.83	.90	.86	270	.61	.62	.61	.67	.68	.67	722
PunDerstand_DeBERTaSampled	.68	.68	.90	.78	91	.44	.57	.50	49	.41	.43	.42	106	.66	.42	.51	147	.52	.71	.60	59	.91	.85	.88	270	.60	.65	.62	.69	.68	.68	722
Arampatzis_DistilBertTokenizer	.68	.72	.84	.78	91	.52	.31	.38	49	.42	.35	.38	106	.64	.56	.59	147	.57	.54	.56	59	.79	.92	.85	270	.61	.58	.59	.66	.68	.66	722
DadJokers_bert_base_uncased	.67	.71	.79	.75	91	.46	.35	.40	49	.42	.39	.40	106	.58	.59	.59	147	.59	.64	.62	59	.85	.86	.85	270	.60	.60	.60	.67	.67	.67	722
ORPAILLEUR_mistral-7b-low	.67	.67	.84	.74	91	.61	.47	.53	49	.53	.22	.31	106	.49	.82	.61	147	.71	.51	.59	59	.89	.79	.84	270	.65	.61	.60	.69	.67	.66	722
NLPalma_BERTd	.67	.63	.82	.71	91	.46	.35	.40	49	.49	.49	.49	106	.66	.53	.59	147	.50	.59	.54	59	.83	.84	.84	270	.60	.60	.59	.67	.67	.67	722
PunDerstand_GuidedAnnotation	.67	.75	.60	.67	5	.57	.67	.62	6	.50	.14	.22	7	.27	.75	.40	4	1	.82	.90	11	.83	.83	.83	12	.65	.64	.61	.73	.67	.67	45
CodeRangers_bert_uncased	.67	.69	.84	.76	91	.40	.51	.45	49	.41	.28	.34	106	.62	.57	.59	147	.57	.61	.59	59	.83	.85	.84	270	.59	.61	.59	.66	.67	.66	722
Arampatzis_Roberta	.66	.72	.78	.75	91	.47	.31	.37	49	.43	.21	.28	106	.54	.62	.58	147	.48	.63	.54	59	.83	.90	.87	270	.58	.57	.56	.64	.66	.65	722
Demonteam_BERTM	.66	.68	.76	.72	91	.48	.45	.46	49	.44	.28	.34	106	.57	.57	.57	147	.52	.56	.54	59	.81	.89	.85	270	.58	.58	.58	.65	.66	.65	722
Arampatzis_MLP	.66	.69	.85	.76	91	.44	.31	.36	49	.42	.23	.29	106	.52	.70	.60	147	.58	.53	.55	59	.85	.84	.85	270	.58	.57	.57	.65	.66	.65	722
Arampatzis_XLNet	.65	.73	.76	.75	91	.45	.35	.39	49	.41	.28	.33	106	.62	.54	.58	147	.46	.61	.53	59	.77	.88	.82	270	.57	.57	.57	.63	.65	.64	722
Arampatzis_Albert	.65	.76	.78	.77	91	.39	.61	.48	49	.40	.27	.32	106	.48	.53	.50	147	.56	.54	.55	59	.87	.84	.85	270	.58	.60	.58	.65	.65	.64	722
UAms_BERT_ft	.63	.68	.85	.75	91	.31	.65	.42	49	.60	.03	.05	106	.55	.64	.59	147	.45	.53	.48	59	.85	.81	.83	270	.57	.58	.52	.66	.63	.60	722
VayamSolveKurmaha_BERT	.60	.58	.38	.46	91	.53	.37	.43	49	.41	.18	.25	106	.52	.63	.57	147	.45	.76	.57	59	.74	.84	.78	270	.54	.53	.51	.59	.60	.58	722

(continued)

Table 3. (*continued*)

run_id	Acc	SD				WS				EX				IR				SC				AID				macro avg			weighted avg			
		P	R	F_1	S	P	R	F_1	S	P	R	F_1	S	P	R	F_1	S	P	R	F_1	S	P	R	F_1	S	P	R	F_1	P	R	F_1	S
NLPalma_PREDCNN	.60	.66	.64	.65	91	.46	.22	.30	49	.44	.36	.40	106	.52	.53	.53	147	.49	.44	.46	59	.69	.83	.75	270	.54	.50	.51	.58	.60	.59	722
Arampatzis_electra	.60	.74	.69	.72	91	0	0	0	49	0	0	0	106	.39	.88	.54	147	0	0	0	59	.78	.89	.83	270	.32	.41	.35	.47	.60	.51	722
NaiveNeuron_fastText	.59	.63	.75	.68	91	.37	.37	.37	49	.46	.37	.41	106	.45	.48	.46	147	.39	.31	.34	59	.76	.79	.78	270	.51	.51	.51	.58	.59	.58	722
VayamSolveKurmaha_BERT	.57	.57	.57	.57	91	.44	.41	.43	49	.36	.20	.26	106	.48	.56	.52	147	.34	.69	.46	59	.83	.73	.78	270	.50	.53	.50	.59	.57	.57	722
DadJokers_RandForest_MLP	.56	.59	.62	.60	91	.41	.35	.38	49	.31	.10	.15	106	.45	.64	.53	147	.47	.44	.46	59	.69	.73	.71	270	.49	.48	.47	.54	.56	.53	722
HumourInsights_RandForest	.55	.59	.68	.63	91	.43	.41	.42	49	.50	.12	.20	106	.47	.44	.45	147	.37	.24	.29	59	.61	.83	.70	270	.50	.45	.45	.53	.55	.52	722
RubyAiYoungTeam	.53	.64	.45	.53	91	.57	.16	.25	49	.41	.08	.14	106	.45	.58	.51	147	.52	.20	.29	59	.56	.85	.68	270	.53	.39	.40	.52	.53	.48	722
Petra&Regina_LogRegr	.53	.64	.45	.53	91	.57	.16	.25	49	.41	.08	.14	106	.45	.58	.51	147	.52	.20	.29	59	.56	.85	.68	270	.53	.39	.40	.52	.53	.48	722
Dajana&Kathy_LogRegr	.53	.64	.45	.53	91	.57	.16	.25	49	.41	.08	.14	106	.45	.58	.51	147	.52	.20	.29	59	.56	.85	.68	270	.53	.39	.40	.52	.53	.48	722
Frane_LogRegr	.53	.64	.45	.53	91	.57	.16	.25	49	.41	.08	.14	106	.45	.58	.51	147	.52	.20	.29	59	.56	.85	.68	270	.53	.39	.40	.52	.53	.48	722
NaiveNeuron_llama3:70b_rag	.53	.41	.49	.45	91	.33	.59	.42	49	.34	.46	.39	106	.59	.35	.44	147	.58	.63	.60	59	.74	.64	.69	270	.50	.53	.50	.57	.53	.54	722
NaiveNeuron_llama3:70b_rag-uae	.53	.41	.52	.46	91	.37	.65	.47	49	.35	.50	.41	106	.59	.33	.43	147	.57	.58	.57	59	.74	.61	.67	270	.50	.53	.50	.57	.53	.53	722
DadJokers_RandForest	.52	.66	.49	.57	91	.46	.24	.32	49	.50	.03	.05	106	.46	.48	.47	147	.60	.10	.17	59	.52	.90	.66	270	.54	.37	.37	.53	.52	.46	722
Tomislav&Rowan_SVM	.51	.54	.29	.37	91	.46	.33	.38	49	.33	.12	.18	106	.42	.51	.46	147	.28	.15	.20	59	.59	.84	.69	270	.44	.37	.38	.48	.51	.47	722
Tomislav&Rowan_LogRegr	.48	.58	.23	.33	91	.33	.12	.18	49	.27	.06	.09	106	.42	.41	.41	147	.44	.12	.19	59	.51	.91	.65	270	.42	.31	.31	.45	.48	.41	722
AB&DPV_MLP3000params	.48	.49	.49	.49	91	.39	.24	.30	49	.30	.07	.11	106	.42	.45	.43	147	.33	.27	.30	59	.53	.73	.62	270	.41	.38	.38	.45	.48	.44	722
PunDerstand_GPT4oFewShot	.47	.22	.27	.24	91	.26	.29	.27	49	.30	.42	.35	106	.67	.31	.42	147	.44	.86	.58	59	.71	.59	.64	270	.43	.46	.42	.53	.47	.47	722
Tomislav&Rowan_NaiveBayes	.44	.90	.10	.18	91	.00	.00	.00	49	.00	.00	.00	106	.40	.31	.35	147	.00	.00	.00	59	.44	.96	.60	270	.29	.23	.19	.36	.44	.32	722
AB&DPV_RandForest250	.38	.63	.11	.19	91	.18	.06	.09	49	.40	.04	.07	106	.26	.17	.20	147	.29	.03	.06	59	.40	.86	.55	270	.36	.21	.19	.38	.38	.29	722
AB&DPV_RandForest500	.38	.57	.09	.15	91	.27	.08	.13	49	.30	.03	.05	106	.27	.18	.22	147	.25	.02	.03	59	.40	.86	.54	270	.34	.21	.19	.36	.38	.29	722
AB&DPV_MLP2000	.37	.00	.00	.00	91	.17	.02	.04	49	.00	.00	.00	106	.00	.00	.00	147	.00	.00	.00	59	.38	.99	.54	270	.09	.17	.10	.15	.37	.21	722
AB&DPV_MLP3000	.37	.00	.00	.00	91	.17	.02	.04	49	.00	.00	.00	106	.00	.00	.00	147	.00	.00	.00	59	.38	.99	.54	270	.09	.17	.10	.15	.37	.21	722
CYUT_GPT-4	.36	.09	.09	.09	91	.23	.69	.35	49	.41	.21	.28	106	.63	.12	.20	147	.49	.68	.57	59	.47	.64	.54	139	.39	.40	.34	.42	.36	.32	591
AB&DPV_DecisionTreeClassifier	.29	.21	.24	.22	91	.17	.10	.13	49	.22	.11	.15	106	.24	.29	.26	147	.12	.17	.14	59	.43	.42	.42	270	.23	.22	.22	.29	.29	.28	722
AB&DPV_GaussianNB	.27	.00	.00	.00	91	.17	.27	.21	49	.23	.07	.10	106	.17	.03	.05	147	.10	.56	.17	59	.53	.50	.52	270	.20	.24	.17	.29	.27	.25	722
CYUT_roBERTa-fine-tuning	.19	.00	.00	.00	91	.03	.06	.04	49	.03	.08	.04	106	.59	.54	.56	147	.51	.73	.60	59	.00	.00	.00	270	.19	.24	.21	.17	.19	.17	722

4.2 Data

The data is an extension of the JOKER parallel wordplay corpus [10]. The train data for Task 3 consists of 1,405 English wordplay instances, with a total of 5,838 professional human French translations. The test filed shared with participants consists of 4,501 English wordplay instances. Over these English puns we used new 376 distinct source texts with 832 corresponding references created by professional French native speaker translators. The maximal number of references per English pun is eight. However, the majority of source text have a single reference. The histogram of the test references per English pun is given in Fig. 1. An example of the source data is:

```
{"id_en":"en_1007",
 "text_en":"Save the whales, spouted Tom."},
```

The corresponding human reference translations are:

```
{"text_fr":"\"Il faut sauver les baleines\", jeta Tom avant de se
↪  tasser."},
{"text_fr":"\"Il faut sauver les baleines\", interjeta Tom."},
{"text_fr":"Moi je sauve les baleines, Tom s'en venta."},
{"text_fr":"Louis évent-a le projet de sauvetage des baleines."},
{"text_fr":"\"Sauvez les baleines\", proclama Tom à tout évent."},
{"text_fr":"\"Sauvez les baleines, cracha Toto, Cétacé!\""}
```

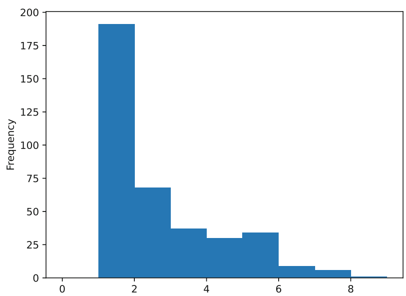


Fig. 1. Histogram of translation references in French per English pun

4.3 Evaluation

We evaluated the runs with the traditional machine translation metrics:

– **BLEU** (BiLingual Evaluation Understudy), which measures the vocabulary overlap between the candidate translation and a reference translation [28]. We used the sacreBLEU implementation[3] with the default tokenizer *13a* which mimics the mteval-v13a script from Moses [29]. We report the BLEU score (harmonic mean) and the BLEU precisions for n-grams on 376 distinct English texts with corresponding 832 reference translations to French.

[3] https://github.com/mjpost/sacrebleu/.

– **BERT Score** from the python *bert-score* package[4] [37]. We report mean values of BERT score precision, recall, and F1 over all 832 references.

4.4 Participants' Approaches

The **Petra&Regina** team [8] submitted a single run. The authors relied on the EasyNMT library, which they used with the Helsinki-NLP Opus-MT model.

The **Frane** team submitted one run. They used neural machine translation models like MarianMT. The translations were refined with a custom module to preserve the pun elements. This module used bilingual dictionaries and contextual embeddings. A similar approach was taken by the **Dajana&Kathy** team.

The **AB&DPV** team [35] used simple prompts with Llama-2-7b and reported that in a number of instances the translations were found to be incomplete or mixing two languages. They submitted a single run.

The **Tomislav&Rowan** team [24] preprocessed the data and used it to build prompts to translate the jokes with the translation pre-trained model (Helsinki-NLP Opus) through the MarianMT framework. The authors judged that EasyNMT was less effective for this task. Two runs were submitted.

The **Farhan** [6] team provided two runs. They used single shot prompting techniques with GPT-4 and GPT-4o.

The **Arampatzis** team submitted six runs for this task employing among others MarianMT, Google Translate, Helsinki-NLP Opus, mBART.

The **Dajana&Kathy** team submitted one run. The provided approach employed sequence-to-sequence models with attention mechanisms, such as Transformer models. In particular, MarianMT and mBART were trained on the pun translation dataset and the pun detection module was implemented to identify puns in the source text. They employed various strategies to preserve the wordplay in the target language such as substituting equivalent puns in French or creatively adapting the humor. Then manual and automatic evaluation was done to ensure the translation quality.

The **UAms** team [5] submitted two runs. MarianMT – a sequence-to-sequence (`Seq2Seq`) model based on the Marian framework was used. The second run was based on the T5 (`t5-base`) model with the same standard preprocessing as for the first run.

The **Jokester** [2] team submitted one run. They also used the MarianMT framework.

Finally, the **Olga** team submitted three runs. The team explored the topic of translating humor from English to Spanish, comparing the BLOOM model with Google Translate. For the BLOOM translations, two different prompts were employed. We do not provide an evaluation of her runs here as we have not done it for Spanish this year.

[4] https://pypi.org/project/bert-score/.

Table 4. Task 3 participants results in terms of the BLEU scores, BLEU n-gram precisions, BERT Score precision, recall, and F1

run_id	BLEU						BERT_Score			
	count	Score	n_1	n_2	n_3	n_4	count	P	R	F1
Arampatzis_GoogleTranslate	376	65.23	78.96	67.48	61.59	57.52	832	91.93	91.82	91.85
Frane_TranslationModel	92	57.13	64.33	58.41	54.66	51.85	279	92.06	91.53	91.77
Dajana&Kathy	376	58.45	71.94	60.27	54.11	49.73	832	91.35	91.00	91.15
UBO_SDL	312	13.17	71.90	57.17	49.13	43.24	598	90.13	90.21	90.15
Tomislav&Rowan_MarianMT	376	58.85	77.11	63.66	56.06	50.45	832	90.82	89.19	89.95
Arampatzis_MarianMT	376	58.85	77.11	63.66	56.06	50.45	832	90.82	89.19	89.95
UBO_ChatGPT	312	13.09	69.90	54.08	46.07	40.31	598	89.12	89.34	89.21
UBO_DeepL	312	11.97	68.53	50.32	41.38	35.11	598	89.06	89.31	89.16
UAms_T5-base_ft	376	48.74	71.75	54.57	45.18	38.05	832	89.53	88.52	89.00
Arampatzis_mBART	376	48.71	70.95	54.40	45.29	38.67	832	88.95	87.41	88.13
Arampatzis_M2M100	376	42.37	68.46	48.73	37.72	29.93	832	88.23	87.23	87.70
UAms_Marian_ft	376	25.69	47.05	28.47	20.74	15.69	832	81.06	82.53	81.74
Farhan_2	376	14.33	23.68	15.84	12.05	9.32	832	69.38	77.14	72.96
Farhan_1	376	9.21	15.92	9.97	7.65	5.92	832	64.30	73.18	68.41
jokester_MarianMT	49	0.29	15.34	0.14	0.08	0.04	112	67.30	66.38	66.80
Arampatzis_opus_mt	63	0.29	15.04	0.23	0.06	0.03	157	66.98	66.05	66.47
Arampatzis_T5	63	0.32	11.35	0.17	0.10	0.06	157	65.91	64.79	65.31

4.5 Results

Eleven teams submitted 23 runs for Task 3. Table 4 shows the results of the CLEF 2024 JOKER track's Task 3. We report the participants results in terms of the BLEU score, BLEU n-gram precisions over the set of 376 English source puns with corresponding 832 references. The BERT Score precisions, recalls, and F1 are averaged over 832 French references.

We make the following observations. First, the best results were obtained by participants who used the commercial machine translation engines such as Google Translate and DeepL integrated into the SDL studio. The MarianMT models, which is similar to BART, showed very similar results. Second, the same models fine-tunes by difference teams are scored differently. Third, the BLEU scores of the UBO submission are very low while the BERT scores are very high. More analysis is needed to investigate this difference.

5 Discussion and Conclusions

This paper has outlined the setup of the CLEF 2024 JOKER Lab, which features shared tasks: one on humour-aware information retrieval, one on the classification of short humorous texts into different humor types, and finally the task of humour translation (English-French). We briefly described the submitted runs to the three tasks and overviewed the results. 103 runs were submitted to the JOKER track with Task 2 on humor classification being the most popular with 54 runs from 18 teams. Task 1 received 26 runs by 10 teams, while 23 runs were sent for Task 3 (11 teams).

For humour-aware IR, we observe that pseudo-relevance feedback RM3 query expansion outperforms BM25 baselines. Cross-encoder rerankers do not perform better than baseline models. However, filtering trained on wordplay detection significantly improves system results. Overall, both precision and recall are very low. Low precision is due to query terms appearing in non-humorous texts, considered topically relevant by retrieval systems. Low recall is likely because the query terms are often absent in both humorous and topically relevant texts, possibly due to text length.

For the humour classification task, the state-of-the-art LLM-based models clearly outperform traditional classifiers. However, the overall scores of all participants are still less than 80% in terms of accuracy as well as the F_1 score per class. These results indicate that the pragmatic is still challenging for LLMs despite their significant recent advances.

We also observe that the wordplay translation task is still challenging for LLMs. The commercial machine translation engines have the highest BLEU overlap with the manual references as well as the BERT score. However, we observe opposite tendencies for these two metrics. Further analysis is needed.

Overall, we introduced a new task of humour-aware information retrieval, we provided a new dataset for humour classification task and an updated parallel corpus of wordplay translation from English to French. In future, we plan to extend our Task 1 to cross-lingual humour-aware information retrieval and to develop specific metrics to evaluate wordplay translation.

Acknowledgments. This project has received a government grant managed by the National Research Agency under the program "Investissements d'avenir" integrated into France 2030, with the Reference ANR-19-GURE-0001. This track would not have been possible without the great support of numerous individuals. We want to thank in particular the colleagues and the students who participated in data construction and evaluation, in particular the students of the Université de Bretagne Occidentale. Please visit the JOKER website for more details on the track (https://joker-project.com/).

References

1. Abu Farha, I., Oprea, S.V., Wilson, S., Magdy, W.: SemEval-2022 task 6: iSarcasmEval, intended sarcasm detection in English and Arabic. In: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pp. 802–814. Association for Computational Linguistics, July 2022. https://doi.org/10.18653/v1/2022.semeval-1.111
2. Baguian, H., Ashley, H.N.: JOKER Track @ CLEF 2024: the Jokesters' approaches for retrieving, classifying, and translating wordplay. In: Faggioli, G., Ferro, N., Galuscakova, P., García Seco de Herrera, A. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)

3. Balaji, S.S., et al.: Vayam Solve Kurmaha @ CLEF 2024: task 2: humor classification according to genre and technique using BERT embeddings and transformers. In: Faggioli, G., Ferro, N., Galuscakova, P., García Seco de Herrera, A. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
4. Bogireddy, N.R., Suresh, S., Rai, S.: I'm out of breath from laughing! I think? A dataset of COVID-19 humor and its toxic variants. In: Companion Proceedings of the ACM Web Conference 2023, pp. 1004–1013. Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3543873.3587591
5. Buijs, L., Cazemier, M., Schuurman, E., Kamps, J.: University of Amsterdam at the CLEF 2024 joker track. In: Faggioli, G., Ferro, N., Galuscakova, P., García Seco de Herrera, A. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
6. Dhanani, F., Abbas, R.: Translating English puns to French. In: Faggioli, G., Ferro, N., Galuscakova, P., García Seco de Herrera, A. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
7. Dsilva, R.R., Bhardwaj, N.: CLEF JOKER 2024 task 2: who's laughing now? Humor classification by genre and technique. In: Faggioli, G., Ferro, N., Galuscakova, P., García Seco de Herrera, A. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
8. Elagina, R., Vučić, P.: Convergential approach in machine learning for effective humour analysis and translation. In: Faggioli, G., Ferro, N., Galuscakova, P., García Seco de Herrera, A. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
9. Epron, P., Guibon, G., Couceiro, M.: CLEF 2024 JOKER task 2: leveraging large language models for humor genre classification. In: Faggioli, G., Ferro, N., Galuscakova, P., García Seco de Herrera, A. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
10. Ermakova, L., Bosser, A.G., Jatowt, A., Miller, T.: The JOKER Corpus: English–French parallel data for multilingual wordplay recognition. In: SIGIR 2023: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2796–2806. Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3539618.3591885
11. Ermakova, L., Miller, T., Bosser, A.G., Palma Preciado, V.M., Sidorov, G., Jatowt, A.: Overview of JOKER 2023 automatic wordplay analysis task 1 – pun detection. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, vol. 3497, pp. 1785–1803, October 2023
12. Ermakova, L., Miller, T., Bosser, A.G., Palma Preciado, V.M., Sidorov, G., Jatowt, A.: Overview of JOKER 2023 automatic wordplay analysis task 2 – pun location and interpretation. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, vol. 3497, pp. 1804–1817, October 2023
13. Ermakova, L., Miller, T., Bosser, A.G., Palma Preciado, V.M., Sidorov, G., Jatowt, A.: Overview of JOKER 2023 automatic wordplay analysis task 3 – pun translation. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, vol. 3497, pp. 1818–1827, October 2023
14. Ermakova, L., Miller, T., Bosser, A.G., Palma Preciado, V.M., Sidorov, G., Jatowt, A.: Overview of JOKER - CLEF-2023 track on automatic wordplay analysis. In:

Arampatzis, A., et al. (eds.) CLEF 2023. LNCS, vol. 14163, pp. 397–415. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-42448-9_26
15. Ermakova, L., Miller, T., Bosser, A.G., Palma Preciado, V.M., Sidorov, G., Jatowt, A.: Science for fun: the CLEF 2023 JOKER track on automatic wordplay analysis. In: Kamps, J., et al. (eds.) ECIR 2023. LNCS, vol. 13982, pp. 546–556. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-28241-6_63
16. Ermakova, L., et al.: CLEF workshop JOKER: automatic wordplay and humour translation. In: Hagen, M., et al. (eds.) ECIR 2022. LNCS, vol. 13186, pp. 355–363. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-99739-7_45
17. Ermakova, L., et al.: Overview of JOKER@CLEF 2022: automatic wordplay and humour translation workshop. In: Barrón-Cedeño, A., et al. (eds.) CLEF 2022. LNCS, vol. 13390, pp. 447–469. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-13643-6_27
18. Ermakova, L., et al.: Overview of the CLEF 2024 JOKER task 1: humour-aware information retrieval. In: Faggioli, G., et al. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
19. Ermakova, L., et al.: Overview of the CLEF 2024 JOKER task 3: translate puns from English to French. In: Faggioli, G., et al. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
20. Frenda, S., et al.: EPIC: multi-perspective annotation of a corpus of irony. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 13844–13857. Association for Computational Linguistics, July 2023. https://doi.org/10.18653/v1/2023.acl-long.774
21. Gepalova, A., Chifu, A.G., Fournier, S.: CLEF 2024 JOKER task 1: exploring pun detection using the T5 transformer model. In: Faggioli, G., Ferro, N., Galuscakova, P., García Seco de Herrera, A. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
22. Kováčiková, J.V., Šuppa, M.: Thinking, fast and slow: from the speed of FastText to the depth of retrieval augmented large language models for humour classification. In: Faggioli, G., Ferro, N., Galuscakova, P., García Seco de Herrera, A. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
23. Macdonald, C., Tonellotto, N.: Declarative experimentation in information retrieval using PyTerrier. In: Proceedings of ICTIR 2020 (2020)
24. Mann, R., Mikulandric, T.: CLEF 2024 JOKER tasks 1–3: humour identification and classification. In: Faggioli, G., Ferro, N., Galuscakova, P., García Seco de Herrera, A. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
25. Narayanan, S., et al.: CLEF 2024 JOKER task 3 : using RoBERTa and BERT-uncased for humour classification according to genre and technique. In: Faggioli, G., Ferro, N., Galuscakova, P., García Seco de Herrera, A. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
26. Palma Preciado, V.M., Palma Preciado, C., Sidorov, G.: NLPalma Joker 2024: yet, no humor with humorousness – task 2 humour classification according to genre and technique. In: Faggioli, G., Ferro, N., Galuscakova, P., García Seco de Herrera, A. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)

27. Palma-Preciado, V.M., et al.: Overview of the CLEF 2024 JOKER task 2: humour classification according to genre and technique. In: Faggioli, G., et al. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
28. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002). https://doi.org/10.3115/1073083.1073135. https://www.aclweb.org/anthology/P02-1040
29. Post, M.: A call for clarity in reporting BLEU scores. In: Proceedings of the Third Conference on Machine Translation: Research Papers, Belgium, Brussels, pp. 186–191. Association for Computational Linguistics, October 2018. https://www.aclweb.org/anthology/W18-6319
30. Saipranav, M., Sridharan, J., et al.: CLEF 2024 JOKER task 3: using BERT and random forest classifier for humor classification according to genre and technique. In: Faggioli, G., Ferro, N., Galuscakova, P., García Seco de Herrera, A. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
31. Shapira, N., Kalinsky, O., Libov, A., Shani, C., Tolmach, S.: Evaluating humorous response generation to playful shopping requests. In: Kamps, J., et al. (eds.) ECIR 2023. LNCS, vol. 13981, pp. 617–626. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-28238-6_53
32. Subramanian, R., Sivaraman, V., et al.: Humour insights at JOKER 2024 task 2: humour classification according to genre and technique. In: Faggioli, G., Ferro, N., Galuscakova, P., García Seco de Herrera, A. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
33. Touvron, H., et al.: Llama 2: open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
34. Van Gysel, C., de Rijke, M.: Pytrec_eval: an extremely fast python interface to trec_eval. In: SIGIR. ACM (2018)
35. Varadi, D.P., Bartulović, A.: JOKER 2024 by AB&DPV: from 'LOL' to 'MDR' using AI models to retrieve and translate puns. In: Faggioli, G., Ferro, N., Galuscakova, P., García Seco de Herrera, A. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
36. Wu, S.H., Huang, Y.F., Lau, T.Y.: Humour classification by fine-tuning LLMs: CYUT at CLEF 2024 JOKER Lab subtask humour classification according to genre and technique. In: Faggioli, G., Ferro, N., Galuscakova, P., García Seco de Herrera, A. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
37. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTscore: evaluating text generation with BERT. In: International Conference on Learning Representations (2020). https://openreview.net/forum?id=SkeHuCVFDr

Overview of LifeCLEF 2024: Challenges on Species Distribution Prediction and Identification

Alexis Joly[1(✉)], Lukáš Picek[1,8], Stefan Kahl[6,11], Hervé Goëau[2], Vincent Espitalier[2], Christophe Botella[1], Diego Marcos[1], Joaquim Estopinan[1,12], Cesar Leblanc[1], Théo Larcher[1], Milan Šulc[10], Marek Hrúz[8], Maximilien Servajean[7], Hervé Glotin[3], Robert Planqué[4], Willem-Pier Vellinga[4], Holger Klinck[6], Tom Denton[9], Ivan Eggel[5], Pierre Bonnet[2], and Henning Müller[5]

[1] Inria, LIRMM, Univ Montpellier, CNRS, Montpellier, France
alexis.joly@inria.fr
[2] CIRAD, UMR AMAP, Montpellier, Occitanie, France
[3] Univ. Toulon, Aix Marseille Univ., CNRS, LIS, DYNI Team, Marseille, France
[4] Xeno-canto Foundation, Hague, The Netherlands
[5] Informatics Institute, HES-SO Valais, Sierre, Switzerland
[6] Cornell Lab of Ornithology, Cornell University, Ithaca, NY, USA
[7] LIRMM, Université Paul Valéry Montpellier, Montpellier, France
[8] Department of Cybernetics, FAV, University of West Bohemia, Pilsen, Czechia
[9] Google Research, San Francisco, USA
[10] Second Foundation, Prague, Czech Republic
[11] Chemnitz University of Technology, Chemnitz, Germany
[12] Laboratoire d'Ecologie Alpine, Grenoble, France

Abstract. Biodiversity monitoring using machine learning and AI-based approaches is becoming increasingly popular. It allows for providing detailed information on species distribution and ecosystem health at a large scale and contributes to informed decision-making on environmental protection. Species identification based on images and sounds, in particular, is invaluable for facilitating biodiversity monitoring efforts and enabling prompt conservation actions to protect threatened and endangered species. The multiplicity of methods developed, however, makes it important to evaluate their performance on realistic datasets and using standardized evaluation protocols. The LifeCLEF lab has been setting up such evaluations since 2011, encouraging machine learning researchers to work on this topic and promoting the adoption of the technologies developed by stakeholders. The 2024 edition proposes five data-oriented challenges related to the identification and prediction of biodiversity: (i) BirdCLEF: bird call identification in soundscapes, (ii) FungiCLEF: revisiting fungi species recognition beyond 0-1 cost, (iii) GeoLifeCLEF: remote sensing based prediction of species, (iv) PlantCLEF: Multi-species identification in vegetation plot images, and (v) SnakeCLEF: revisiting snake species identification in medically important scenarios. This paper overviews the motivation, methodology, and main outcomes of those five challenges.

L. Goeuriot et al. (Eds.): CLEF 2024, LNCS 14959, pp. 183–207, 2024.
https://doi.org/10.1007/978-3-031-71908-0_9

1 LifeCLEF Lab Overview

Accurately identifying organisms observed in the wild is an essential step in ecological studies. It forms the foundation for understanding species interactions, population dynamics, and ecological processes, allowing researchers to accurately assess biodiversity, track changes over time, and make informed management and conservation decisions. However, observing and identifying living organisms requires high levels of expertise. For instance, vascular plants alone account for more than 300,000 different species and the distinctions between them can be quite subtle. The worldwide shortage of trained taxonomists and curators capable of identifying organisms has come to be known as the *taxonomic impediment*. Since the Rio Conference of 1992, it has been recognized as one of the major obstacles to the global implementation of the Convention on Biological Diversity. In 2004, Gaston and O'Neill [20] discussed the potential of automated approaches for species identification. They suggested that if the scientific community were able to (i) produce large training datasets, (ii) precisely evaluate error rates, (iii) scale-up automated approaches, and (iv) detect novel species, then it would be possible to develop a generic automated species identification system that would open up new vistas for research in biology and related fields.

Since the publication of [20], automated species identification has been widely studied [14,40,58,65,69] and is now a key technology in most citizen science monitoring apps, e.g., iNaturalist, eBird and Pl@ntNet [6]. Nevertheless, the development of new approaches continues to expand rapidly, in particular for processing new types of data such as passive sensors, camera traps, or autonomous vehicles [16,70,76]. Biodiversity monitoring through AI approaches is now recognized as a key solution to collect and analyze vast amounts of data from various sources, enabling us to gain a comprehensive understanding of species distribution, abundance, and ecosystem health [2,5]. This information is essential for making informed conservation decisions and identifying areas needing protection.

To measure progress of AI-assisted biodiversity monitoring in a sustainable and repeatable way, the LifeCLEF virtual lab was created in 2014 as a continuation and extension of the plant identification task that had been run within the ImageCLEF lab since 2011 [23–25]. Since 2014, LifeCLEF has expanded the challenge by considering animals and fungi in addition to plants and including audio and video content in addition to images [30–39]. Nearly a thousand researchers and data scientists participate yearly to LifeCLEF to analyze the data, submit predictions and benefit from the shared evaluation tools. The aim of this paper is to present the synthesis of the 2024th edition of LifeCLEF, which comprises five challenges synthesized in Table 1.

The systems used to run the challenges (registration, submission, leaderboard, etc.) were the Kaggle platform for the BirdCLEF and GeoLifeCLEF, and the Hugging Face for the PlantCLEF, SnakeCLEF, and FungiCLEF challenges. Four of the challenges (GeoLifeCLEF, SnakeCLEF, PlantCLEF, and FungiCLEF) were organized jointly with FGVC 11, an annual workshop dedicated to

Table 1. LifeCLEF challenges data overview. The provided datasets vary in modality, size, and complexity as each challenge addresses different aspects of automated species identification.

	Modality	Species	Items	Task	Metric
BirdCLEF	audio	182	25K	Multi-label classification	ROC-AUC
SnakeCLEF	images metadata	1,784	190K	Classification	ad-hoc metric
FungiCLEF	images metadata	4,759	400K	Open-set classification	ad-hoc metric
PlantCLEF	images (SD+HD)	7,806	1.4M	Multi-label classification	Samples F1
GeoLifeCLEF	sat. images time-series tabular	10,358	6.6M	Multi-label classification	Sample-Average F1

Fine-Grained Visual Categorization, held in conjunction with the CVPR international conference on computer vision and pattern recognition.
In total, 1045 data scientists or research teams participated in the LifeCLEF 2024 edition by submitting runs to at least one of the five challenges (966 only for the BirdCLEF challenge). Only some of them managed to get the results right, and 18 of them went all the way through the CLEF process by writing and submitting a *working note* describing their approach and results (for publication in CEUR-WS proceedings. In the following sections, we provide a synthesis of the methodology and main outcomes of each of the five challenges. More details can be found in the extended overview reports of each challenge and in the individual working notes of the participants (references provided below).

2 BirdCLEF Challenge: Bird Call Identification in Soundscapes

A detailed description of the challenge and a more complete discussion of the results can be found in the dedicated working note [41].

2.1 Objective

Birds are vital indicators of biodiversity change due to their mobility and diverse habitat requirements. Changes in bird species assemblage and numbers can signal the success or failure of restoration projects. Traditional observer-based bird surveys over large areas are expensive and logistically challenging. Passive acoustic monitoring (PAM) combined with machine learning enables conservationists to sample larger areas with higher temporal resolution, providing detailed insights into the relationship between restoration efforts and biodiversity.

The Western Ghats, a Global Biodiversity Hotspot along India's southwestern coast, support extraordinary biodiversity across various ecosystems, from high-elevation forest-grassland mosaics to wet-evergreen rainforests. This region also sustains large human populations relying on forest resources. The Western Ghats host a high diversity of bird species, including many endemic and endangered species. However, significant landscape and climatic changes are threatening this biodiversity, highlighting the need for advanced conservation tools to rapidly assess and monitor bird diversity. The competition aims to identify endemic bird species in the Western Ghats' sky-islands using soundscape data, detect and classify endangered bird species with limited training data, and detect and classify poorly understood nocturnal bird species.

2.2 Dataset

We built on the experience from previous editions and adjusted the task to encourage participants to focus on task-specific model designs. We carefully selected training and test data to match this objective. As in previous iterations, Xeno-canto was the primary source for training data, while expertly annotated soundscape recordings were used for testing. We emphasized bird species that are typically underrepresented in large bird sound collections, such as those that are ecologically important but difficult to train a classifier due to their rare or elusive nature. However, we also included common species to allow participants to train effective recognition systems. To find suitable test data, we considered various sources with differing complexities, such as call density, chorus, signal-to-noise ratio, and man-made sounds, as well as quality differences like mono and stereo recordings. This year, we also included unlabeled training data similar to the test data, enabling participants to explore alternative training methods such as self-supervised learning.

2.3 Evaluation Protocol

The challenge was hosted on Kaggle, maintaining an evaluation mode similar to previous iterations with hidden test data and a code competition format. We used a version of macro-averaged ROC-AUC that skips classes without true positive labels as the metric. This approach allowed us to assess system performance independent of fine-tuned confidence thresholds, emphasizing per-species performance rather than per-sample performance. Participants were tasked with identifying species from short audio segments extracted from labeled soundscape data. We used 5-second segments, balancing typical signal length with sufficiently long context windows. The dataset size was kept reasonably small (¡50 GB) and easy to process. Additionally, we provided introductory code repositories and write-ups to lower the entry barrier for the competition.

2.4 Participants and Results

A total of 1,198 participants, organized into 974 teams, participated in the BirdCLEF 2024 challenge, submitting more than 30,000 runs. Figure 1 shows the performance of the top 25 runs. The primary metric was the private leaderboard score, revealed after the submission deadline to prevent probing of the hidden test data. Throughout the competition, participants were able to see their public score, which was calculated based on 35% of the test data.

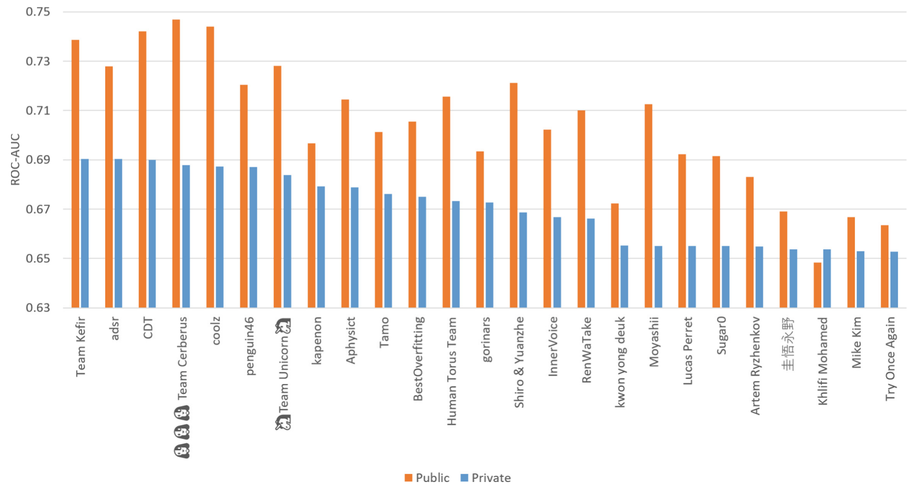


Fig. 1. BirdCLEF 2024 results. Top 25 teams sorted by private leaderboard score.

The baseline score in this year's edition was 0.5 (due to the metric) with random confidence scores for all birds for all segments. The best submission achieved a score of 0.690 (public 0.738) and the top 10 best performing systems were within only 1.5% difference in score. The majority of methods employed ensembles of convolutional neural networks, differing primarily in their pre- and post-processing techniques and the neural network backbones they used. Top participants leveraged unlabeled soundscape data to enhance their scores and adapt to the test data's acoustic domain. Given the restricted CPU runtime for submissions, participants prioritized speeding up model inference and using efficient architectures, with EfficientNet backbones being particularly popular. Additionally, participants explored ONNX and openVINO to further boost model inference speed. More details about the methods employed and the analysis of the results can be found in the detailed report of the task [41] and in the individual working notes of participants.

3 GeoLifeCLEF Challenge: Species Composition Prediction with High Spatial Resolution at Continental Scale Using Remote Sensing

Comprehensive details on the challenge and an extensive discussion of the results are available in the dedicated working note [51].

3.1 Objective

Predicting species presence within specific areas is crucial for ecological research and biodiversity conservation. Accurate predictions support decisions related to protecting endangered species, land use planning, establishing protected zones, and developing sustainable agricultural practices. Nonetheless, species distributions are often influenced by intricate local variables that are difficult to quantify, such as interactions between populations, landscape connectivity, historical habitat conditions, and biases in data collection methods. Traditional ecological models often struggle with these complexities, resulting in predictions with limited spatial resolution. Furthermore, many species are underrepresented due to sampling biases. GeoLifeCLEF addresses these challenges by evaluating models on a vast scale, encompassing thousands of species, achieving spatial resolutions up to 10 m, and leveraging millions of occurrence data points.

3.2 Dataset

The GeoLifeCLEF 2024 dataset contains species observation data, including presence-only occurrences and presence-absence surveys, alongside various environmental predictors. The dataset provides diverse environmental rasters, Sentinel2 satellite images, a 20-year climatic time series, and satellite time-series point values. Following on the work and dataset provided in the previous edition [4], we took most of the already provided Presence-Only (PO) occurrences (5 million) but tripled the Presence-Absence (PA) survey records to 90 thousand. Same as last year, the presence-absence data was split into training and test sets (95/5) using a spatial block hold-out procedure [63] with a spatial grid with 10×10 km cells enabling comprehensive model evaluation. The test cells were randomly selected to ensure balance in biogeographical regions. In addition to the raw data, we have provided all the environmental predictors as pre-extracted scalar values in separated CSV files. Furthermore, the time-series data were provided in a 3d cube format (as torch tensors). More details about the dataset are available in the dedicated working note [51].

3.3 Evaluation Protocol

Same as in the previous edition [51], the evaluation metric was selected as the sample averaged F1 score (F_1). The F_1-score serves as a metric to gauge the degree of agreement between the predicted and actual species composition observed within a specific geographical area and timeframe. In the context of ecological surveys, such as those conducted in Protected Areas (PAs), each survey instance i is associated with a ground-truth set of labels Y_i, representing the plant species identified by experts within a defined grid. Given this setup, and a list of predicted labels $\widehat{Y}_{i,1}, \widehat{Y}_{i,2}, \ldots, \widehat{Y}_{i,R_i}$, the micro F_1-score can be computed as follows:

$$\mathrm{F}_1 = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + (FP_i + FN_i)/2}, \tag{1}$$

$$\text{where} \begin{cases} TP_i = \# \text{ of correctly predicted species, i.e., } |\widehat{Y}_i \cap Y_i|. \\ FP_i = \# \text{ of species predicted but not observed, i.e., } |\widehat{Y}_i \setminus Y_i|. \\ FN_i = \# \text{ of species not predicted but present, i.e., } |Y_i \setminus \widehat{Y}_i|. \end{cases} \tag{2}$$

This formulation encapsulates the precision and recall elements crucial for assessing the accuracy of predictive models in ecological studies.

3.4 Organizer's Baselines

We provide a variety of weak and strong baselines for all participants to allow a good starting point, continual performance increase, and working with different modalities. Considering the significant extent to which this baseline's performance can be enhanced, we encouraged participants to experiment with various techniques, architectures, losses, etc. Below, we briefly describe all baselines:

Naive Baselines. With the dense and numerous observation data, one can naively predict the species' presence by selecting a set of the most common species within administrative or bio-geographical regions. For instance, predicting the top-25 most common species in the PA data results in a sample-averaged F_1 of 11.6%. Using the same approach but with the PO data results in an F_1 of 8.1%, showing a distribution shift between the two types of data.

Small Residual Convolutional Neural Networks for Data Cubes. Starting from a Resnet18 architecture, we have developed an even lighter model adapted to the small input size of GLC's cube data (respectively $19 \times 12 \times 4$ for the climatic time series and $21 \times 4 \times 6$ for the Landsat time series). When trained on the PA data with the Binary Cross Entropy loss (BCE), they achieved a sample-averaged F_1 score of respectively 0.259 and 0.266.

Swin Tranformer for the Sentinel2 Images. We slightly modified the architecture of a Swin-v2-t to allow input of all 4 modalities of Sentinel2 data (RGB+IR) rather than just three. It was also trained with the BCE loss on the PA data but resulted in a lower F_1 score of 0.235.

Multi-modal Model. A multimodal model merging all three individual models mentioned above was developed using an MLP (Multi-Layer Perceptron) for the fusion head. It allows reaching an F_1 score of 0.316, demonstrating the task's inherent multimodality.

3.5 Participants and Results

51 Kaggle registrants participated in the GeoLifeCLEF 2024 challenge with at least one valid submission (submissions duplicated from the organizers's baselines were filtered out). A total of 1184 entries (i.e., *runs*) were submitted with an average of 23 entries per participant and a maximum of 175 for the participant who ranked first on the leaderboard. Details of the methods and systems used by the participants who submitted a working note are synthesized in the overview paper of the task [51] and described in more detail in the participant's working notes [8,9,11,44,47,67]. In Fig. 2, we report the performance achieved by all participant's methods as well as the baseline methods developed by the organizers. Hereafter, we provide a short overview of the methods of the two best teams who submitted a working note (top2, top3, and top5 on the leaderboard):

AI2Lab team (Top2) [8]: This team started from the multi-modal model provided as the baseline by the organizers and made several significant improvements: (i) addition of a fourth modality (i.e., tabular environmental data encoded with an MLP), (ii) use of PO data samples through a pseudo-labeling procedure, (iii) use of an improved encoder for the Sentinel2 images (pre-trained with self-supervised learning on an external dataset), (iv) use of an ensemble of models optimized on different folds and, (v) optimization of the detection threshold. They finally got an F_1 score of 0.368 on the private leaderboard.

Miss Qiu (Top3) [44]: This team also started from the multi-modal model provided as a baseline but used an alternative fusion method based on cross-attention rather than MLP (which slightly improved the performance). They also introduced a number of improvements, some similar to the AI2Lab team (e.g., the use of an ensemble of k-fold models) and some different, such as (i) the enrichment of predictions with species frequent in neighboring PA and PO samples, (ii) the optimization of the number of returned species, or (iii) the use of various data augmentation techniques (including mixup). They finally got an F_1 score of 0.353 on the private leaderboard.

BernIgen (Top5) [11]: This team first worked on a model using only tabular data based on the XGBoost method (known to work very well on classical species distribution models). They have previously reduced the dimensionality of the input data with a PCA (Principal Component Analysis) and also the number of output species by keeping only the most likely species (about 10% of the species). This model alone already delivers pretty good performance (F_1 score of 0.31). They improved prediction performance by adaptively predicting the number of species to return for each test plot using a regression model (also based on XGBoost). This allowed gaining one more point of F_1 score. Finally, they combined this model with the multi-modal model provided by the organizers and got an F_1 score of 0.349 on the private leaderboard.

3.6 Outcomes

The main outcomes we can derive from the GeoLifeCLEF 2024 are the following:

– Provided baselines had a positive impact on overall performance.
– Proactive engagement with the community and continual release of better baselines increases the impact.
– The use of multi-modal models with specific encoders for each modality is the main key to success.
– The provision of more PA data in the training dataset enabled much higher performance than last year's challenge (for which the best F_1 score was 0.27).
– Reciprocally, the use of PO data proves less beneficial, with only minor gains compared to models trained solely on presence/absence data.

For the future, it seems important to understand why improving performance with presence-only data is difficult, even though it is much larger. The presence of observation bias is clearly a plausible reason (some species are observed more than others), but it seems the spatial scale of the test set's plots may also be an issue. They are indeed quite small (10×10 m on average) and do not necessarily reflect the presence of all the species in larger areas such as the one considered by the models. Moreover, the locations of these plots themselves follow specific protocols, which may introduce observation biases different from those of presence-only data.

4 FungiCLEF Challenge: Revisiting Fungi Species Recognition Beyond 0-1 Cost

Comprehensive details on the challenge and an extensive discussion of the results are available in the dedicated working note [55].

4.1 Objective

Efficient and scalable species recognition is crucial for large-scale initiatives like citizen science projects [56,66], which often operate with limited computational resources. In practice, accurate species identification relies on visual observations of the specimen and additional contextual data such as habitat, substrate, GPS coordinates, and temporal factors. This challenge sets a significant benchmark by integrating visual and contextual information, leveraging rich metadata, precise annotations, and standardized baselines available to all participants. Given that mushrooms are frequently foraged for consumption, the competition also addresses scenarios related to edible$\longleftrightarrow$poisonous misclassifying.

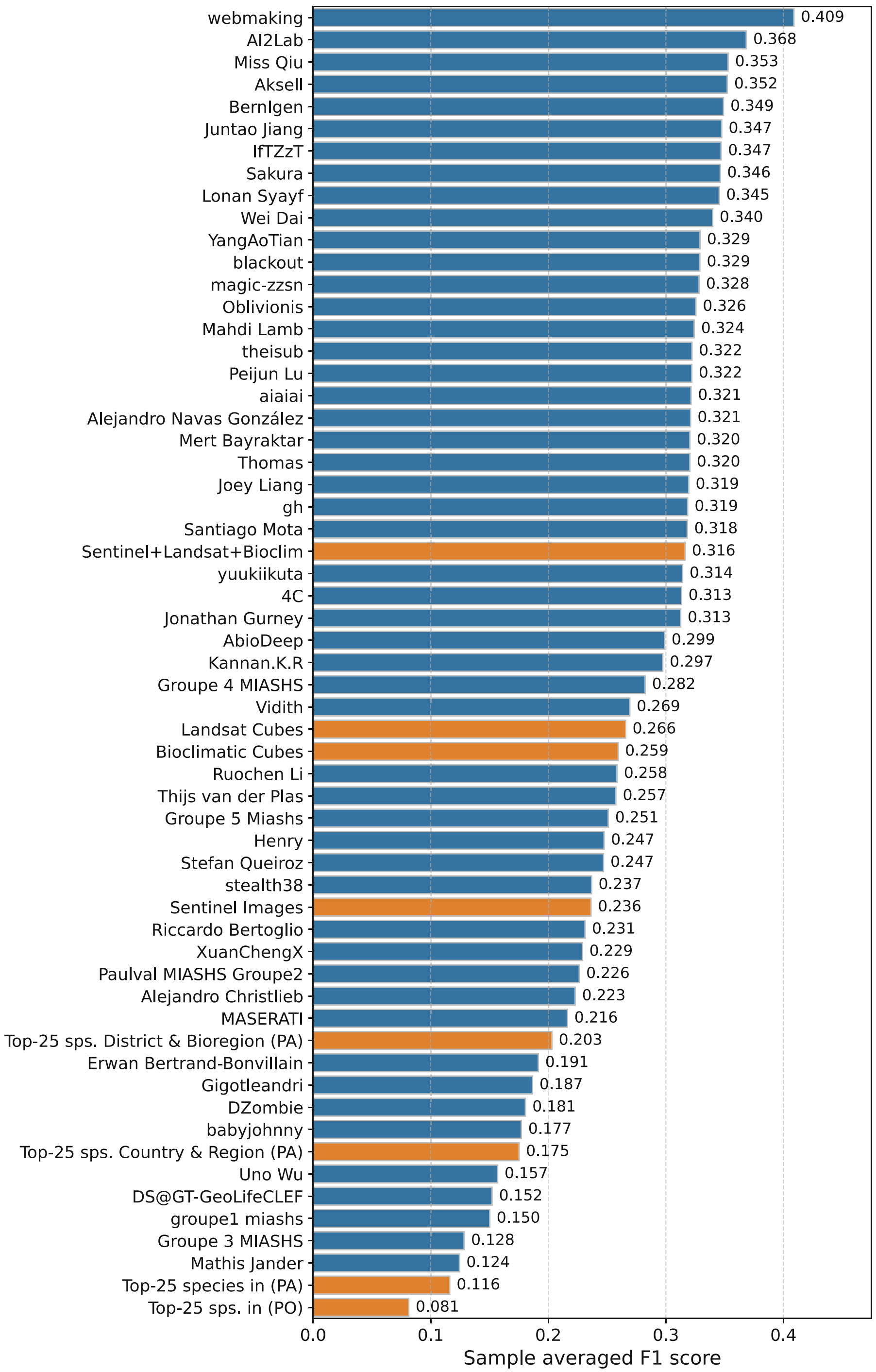


Fig. 2. GeoLifeCLEF 2024 results. All 51 teams. Orange depicts baselines. (Color figure online)

The task requires participants to develop a classification model that generates a ranked list of predicted fungi species for each observation. Each observation includes multiple photographs of the same specimen and geographical location data. The classification model must comply with stringent constraints on memory usage and inference (prediction) time, specifically within a maximum of 120 min, using a dedicated HuggingFace server instance (Nvidia T4, 4 vCPUs, 15 GB RAM, 16 GB VRAM).

Table 2. FungiCLEF 2024 dataset statistics for each subset.

Subset	Species	→ Known/Unknown	Images	Observ.
Training	1,604	1,604 / –	295,938	177,170
Validation	3,299	1,084 / 1,629	91,231	45,021
Test	1,398	749 / 649	41,177	22,412
↪*CzechFungi App*	137	94 / 43	393	215
↪*Atlas of Danish Fungi*	1,261	721 / 540	40,784	22,197

4.2 Dataset

The FungiCLEF 2024 dataset builds upon the previous editions of the FungiCLEF [60,61] and the Danish Fungi 2020 dataset [57]. All the data is derived from a citizen science platform – the Atlas of Danish Fungi. Each fungi observation in this dataset has undergone expert validation, ensuring high-quality species labels. The dataset features rich observation metadata, i.e., information about habitat, substrate, timestamp, location, etc. Provided subsets (i.e., training, validation, and test) are briefly described below, and their statistics in detail are listed in Table 2.

The training set is based on 295,938 training images (177,170 observations) of 1,604 species. The dataset is built exclusively from the Danish Fungi 2020 data by combining the training and public test sets. This results in 295,938 training images across 1,604 species primarily observed in Denmark.

The validation set comprises expert-validated observations with species labels collected solely in 2022. This subset includes around 3,299 fungi species and contains 45,021 observations with many "unknown" species.

The test set is based on two subsets originating from two sources (e.g., Atlas of Danish Fungi and CheckFungi Application) and two countries, e.g., Denmark and Czechia. The CheckFungi is a small subset containing just around 200 submissions and is included primarily as a control set to prevent cheating. The test set was split 80/20 for public and private evaluation, respectively.

4.3 Evaluation Protocol

The task involves developing a classification model to predict species from a given set of real fungi observations accompanied by metadata. This model should adhere to a memory footprint constraint of a maximum of 1GB and prioritize minimizing risks to human safety, mainly by reducing misclassification between poisonous and edible species. The FungiCLEF 2024 challenge employed 2 decision-making scenarios, focusing on minimizing the empirical loss $L = \sum_i W(y_i, q(x_i))$, where $q(x)$ represents the decision rule for observations x, and y denotes the true labels. The cost function $W(y, q(x))$ was tailored for each scenario:

- **Track 1**: Standard classification incorporating an "*unknown*" category;
- **Track 2**: Penalization for edible and poisonous species confusion;
- **Track 3**: A user-centric loss combining Track1 and Track2;

4.4 Participants and Results

Seven teams participated in the FungiCLEF 2024 challenge; of these, six outperformed the baseline with EfficientNet-B1, and five submitted working notes. Details of the best methods and systems used are synthesized in the challenge overview paper [55] and further developed in participants working notes [7,10,18,68,73]. Achieved performance is reported in Fig. 3. This year, the three tracks of FungiCLEF have three different best-performing submissions by three different teams:

The best-performing submission in Track 1 by *Jack Etheredge* [18] combined visual information with metadata using MetaFormer-0 and MetaFormer-2 [15] and further improved the ensemble by a vision-only CAFormer-S18 [75], and proposed a novel application of openGAN [43] for open-set recognition of fine-grained images utilizing WGAN-GP [28].

The best scores in Track 2 were achieved by team *upupup* [68], using Dynamic MLP [74] for the fusion of image features and metadata, identifying unknown classes using an entropy-based approach, training with a marginal expected loss for recognizing poisonous mushrooms while maintaining accuracy.

Finally, the best score for Track 3 was achieved by team *IES* [73], utilizing a Swin Transformer V2 Base [45] for image feature extraction, encoding metadata similarly to the approach of Ren et al. [62] from the previous edition of FungiCLEF, and introducing 1. a poisonous re-ranking that prevents predicting an edible species when there is a significant chance of the sample being poisonous, and 2. a genus loss improves the feature space's regularization.

5 PlantCLEF Challenge: multi-species plant identification in vegetation plot images

A detailed description of the challenge and a more complete discussion of the results can be found in the dedicated working note [26] and the working note participants [12,19,29].

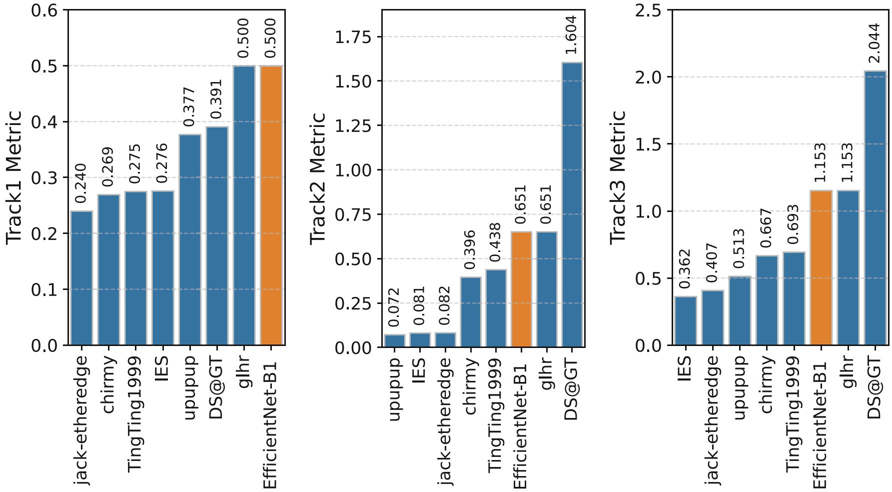


Fig. 3. Private Leaderboard – FungiCLEF 2024 competition – All 7 teams. The orange color depicts baseline performance.

5.1 Objective

Vegetation plot inventories are crucial for ecological studies, enabling standardized sampling, biodiversity assessment, long-term monitoring, and large-scale remote surveys. They provide valuable data on ecosystems, biodiversity conservation, and evidence-based environmental decision-making. Plot images, typically 0.5×0.5 m in size, are meticulously analyzed by botanists who identify all species present. They also quantify species abundance using indicators like biomass, qualification factors, and areas occupied in photographs. AI could greatly improve the efficiency of surveys (with, for example, the participation of non-specialists), thereby increasing the frequency and coverage of ecological studies.

While it is now possible to access very large volumes of images of individual plants and to train very large classification models [21,22], a multi-label declination on large plot images would require complete annotation of all visible species to consider supervised learning of classification models. Unfortunately, such data doesn't exist nowadays and would require considerable efforts to be produced. The PlantCLEF 2024 challenge aims instead to evaluate approaches using classical observations of individual plants as training data, despite the discrepancies between training and test data, as shown in the Fig. 4. Specifically, the challenge is a weakly-supervised multi-label classification task aimed at predicting all plant species visible in high-resolution plot images but with single-label plant images as training data. One of the main difficulties lies in the domain shift between the high-resolution test images of vegetation plots with potentially many species and the training data, which primarily consists of close-up images of individual plants collected through the collaborative platform Pl@ntNet [1].

Furthermore, different weather conditions and shooting angles, along with varying phenological stages, can increase data disparity. Collaborative data might be overrepresented by opportunistic views of flowers, which facilitate identification. In contrast, vegetation plots are typically observed multiple times over one or several years without prior assumptions about the plants' phenological stages (some may be flowering, others fruiting, some in seedling stage, and others senescent or affected by disease).

(a) Vegetative plot (test) (b) Single plant images (training)

Fig. 4. PlantCLEF 2024: illustration of the visual discrepancy between (a) the test set, composed exclusively of vertical top-down views potentially showing many plant species, and (b) the training set, based on images of individual plants, primarily focusing on specific organs (flowers, fruits, leaves, stems).

5.2 Dataset

The training set is composed of observations of individual plants, similar to those used in previous editions of PlantCLEF. More precisely, it is a subset of the Pl@ntNet training data focusing on south western Europe and covering 7,806 plant species. It contains about 1.4 million images extended with some images with trusted labels aggregated from the GBIF platform to complete the less illustrated species. Links to original images are provided in the 'url' column of the metadata csv file. The images have a relatively high resolution (the minimum side is 800 pixels) to allow the use of classification models that can handle relatively large resolution inputs and may reduce the difficulty of predicting small plants in large vegetative plot images. Images are pre-organized into subfolders by class (i.e., by species) and split into a predefined train-validation-test sets to facilitate the training of individual plant classification models.

The test set is a compilation of several image datasets of plots in different floristic contexts, including Pyrenean and Mediterranean floras. These datasets are all produced by experts and consist of a total of 1,695 high-resolution images. The shooting protocol can vary significantly from one context to another: the use of wooden frames or measuring tape to delimit the plot or not, angles of view more or less perpendicular to the ground. Additionally, the quality of the images may vary depending on the weather, which can result in more or less pronounced shadows, blurry areas, etc.

For participants who may have difficulty finding the computational power necessary to train a plant image identification model on such a large volume of data, or to enable direct work with a pre-trained backbone, two pretrained models are shared through Zenodo [27]. Both are based on a vision transformer architecture initially pretrained with the dinov2 self-supervised learning approach [13,49] and fine-tuned on PlantCLEF 2024 training data (with a classical softmax and cross-entropy loss function).

5.3 Evaluation Protocol

The aim of the challenge is to exhaustively list the presence of every plant species on each high-resolution vegetation plot image, from among more than 7,800 species, bearing in mind that plots are generally 50×50 cm in size, and that it's rare for there to be dozens and dozens of species simultaneously.

The metric chosen to differentiate the runs of the participants is the F1 score, adapted to finding a good compromise between recall and precision, i.e. not proposing too many species at the risk of being imprecise, but at the same time not proposing too few species at the risk of being incomplete. Among the several variants of F1 score, the sample-average version is selected as the primary evaluation metric of the challenge (i.e. the average of the F1 scores calculated individually for each vegetation plot). Two other F1 scores variants, namely the micro-average and macro-average, are also shown for information purposes (noticing that the macro-average is difficult to interpret because of missing species in the test set and that the micro-average is known to be biased by data imbalance).

The use of the metadata (image names, EXIF data, licenses) is authorised provided that, for each run using metadata, an equivalent run using only the visual information without metadata in submitted in order to assess the raw contribution of a purely visual analysis. The use of additional data is permitted provided that an equivalent run with only the data provided is submitted to enable more accurate and fair comparisons.

5.4 Participants and Results

The challenge is hosted on the hugging face platform, providing an opportunity for researchers and enthusiasts to contribute to the development of plant recognition in such new context.

Of the 83 teams officially registered on the CLEF registration system for LifeCLEF, 34 registered specifically for the PlantCLEF challenge. On the Hugging Face platform hosting the challenge, 9 teams attempted to submit runs, and in the end 7 teams were able to submit a total of 181 runs. Details of the best methods and systems used are synthesized in the overview working notes paper of the task [26]. In Fig. 5 we report the best performance achieved for each team.

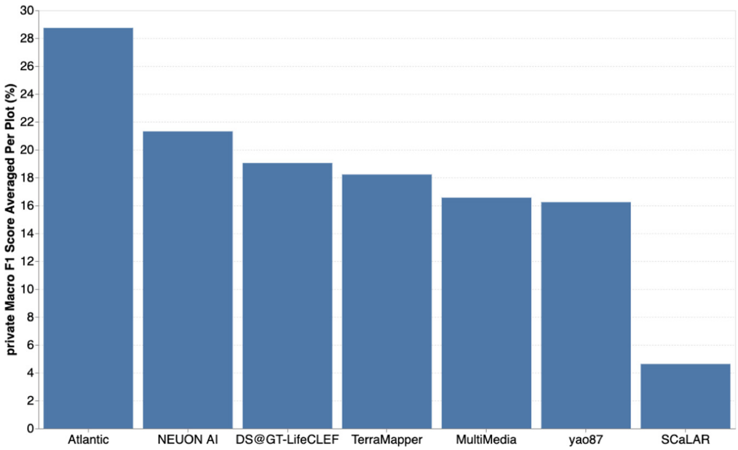


Fig. 5. PlantCLEF 2024: top samples F1 scores for each team.

The main outcomes we can derive from that results are the following:

- Despite the sharing of pre-trained and finetuned state-of-the-art models on a large volume of data specifically on the flora studied, overall performance is low and does not exceed an F1 score of 29%.
- Highest scores were achieved by combining tiling of the high definition images and Vision Transformers models.
- A direct method based on the supplied dinov2 pre-trained model and a tiling approach achieves a F1 score of 22.19% at best, according to participants' working notes.
- The use of an additional background analysis method, based on zero-shot learning with segment-anything [42], proves effective to gain a few extra points, but at the cost of significant computing time.
- Metadata can reveal plots photographed repeatedly over the years, enabling combined predictions for better accuracy. This approach reflects botanists' method of refining identifications through ongoing photo series analysis.

6 SnakeCLEF Challenge: Revisiting Snake Species Identification in Medically Important Scenarios

Comprehensive details on the challenge and an extensive discussion of the results are available in the dedicated working note [53].

6.1 Objective

Given the significant impact of venomous snakebites, creating a robust system to identify snake species from photos is crucial for biodiversity and global health. With over half a million annual deaths and disabilities, understanding the global distribution of 4,000+ snake species through image differentiation enhances epidemiology and treatment outcomes. Despite machines showing accuracy in predictions, especially with long-tailed distributions and 1800 species [3], challenges persist in neglected regions. The next step involves testing in specific tropical and subtropical countries while considering species' medical importance for more reliable machine predictions.

The SnakeCLEF challenge [50,52,54,59] aims to be a major benchmark for observation-based snake species identification. The goal of the task is to create a classification model that returns a ranked list of predicted species for each set of images and location (i.e., snake observation) and minimize the danger to human life and the waste of antivenom if a bite from the snake in the image were treated as coming from the top-ranked prediction. The classification model will have to (i) fit memory footprint limits and a prediction time limit (60 min) within a given HuggingFace server instance (Nvidia T4 small 4vCPU, 15GB RAM, 16GB vRAM), (ii) minimize the danger to human life, i.e., the venomous $\longleftrightarrow$ harmless confusion, (iii) generalize well to all geographic regions.

6.2 Dataset

The training dataset was constructed from observations submitted to the citizen science platforms iNaturalist and HerpMapper and includes around 110,000 real snake observations with community-verified species labels. While constructing the dataset, the species records were sampled based on the country of origin in order to lower the bias towards North America and Europe. Apart from image data, we have provided information about medical importance (i.e., how venomous the species is) and country-species relevance for each snake observation. We list the dataset statistics in Table 3.

6.3 Evaluation Protocol

To motivate research in recognition scenarios with uneven costs for different errors, such as mistaking a venomous snake for a harmless one, we again went beyond the 0-1 cost common in image classification. In addition to Accuracy and macro averaged F_1, we use two metrics (introduced last year) that consider

Table 3. SnakeCLEF 2024 dataset statistics for each subset.

Subset	#Species	#Countries	#Images	#Observations
Training	1,784	212	168,144	95,588
↪*iNaturalist*	*1,784*	*210*	*154,301*	*85,843*
↪*HerpMapper*	*889*	*119*	*13,843*	*9,745*
Validation	1,599	177	14,117	7,816
Private Test	199	12	8,865	4,226
↪*India*	*76*	*1*	*2,892*	*2,395*
↪*Central America*	*107*	*4*	*5,188*	*1,370*
↪*Central Africa*	*80*	*4*	*786*	*462*

venomous $\longleftrightarrow$ harmless confusion and different error costs, i.e., penalizing misclassification of a venomous species with a harmless one more than the other way around. We also calculated two standard metrics, macro averages F1 Score and Accuracy.

The two above-mentioned metrics (T_1 and T_2) are then defined as follows:

$$T_1 = \frac{w_1 \mathrm{F}_1 + w_2 \mathrm{C}_{h \longrightarrow h} + w_3 \mathrm{C}_{h \longrightarrow v} + w_4 \mathrm{C}_{v \longrightarrow v} + w_5 \mathrm{C}_{v \longrightarrow h}}{w_1 + w_2 + w_3 + w_4 + w_5}, \tag{3}$$

where C is equal to 1–ratio of misclassified samples, confusing h-armless and v-enomous species. This metric has a lower bound of 0% and an upper bound of 100%. The lower bound is achieved when all species are misclassified, including misclassifications of harmless species as venomous and vice versa. On the other hand, if the F1-score reaches 100%, indicating the correct classification of all species, each C value must be zero, leading to an overall score of 100%.

$$T_2 = \sum_i L(y_i, \hat{y}_i), \qquad L(y, \hat{y}) = \begin{cases} 0 & \text{if } y = \hat{y} \\ 1 & \text{if } y \neq \hat{y} \text{ and } p(y) = 0 \text{ and } p(\hat{y}) = 0 \\ 2 & \text{if } y \neq \hat{y} \text{ and } p(y) = 0 \text{ and } p(\hat{y}) = 1 \\ 2 & \text{if } y \neq \hat{y} \text{ and } p(y) = 1 \text{ and } p(\hat{y}) = 1 \\ 5 & \text{if } y \neq \hat{y} \text{ and } p(y) = 1 \text{ and } p(\hat{y}) = 0 \end{cases}, \tag{4}$$

where the function p returns 0 if y is a harmless species and 1 if it is venomous.

6.4 Participants and Results

This year, a total of 14 teams participated in the SnakeCLEF. However, just nine teams submitted solutions different from the baseline, and four submitted working notes. Details of the best methods and systems used are synthesized in the competition overview paper [53], with further elaboration available in the individual working notes submitted by the participants [17,48,64,71].

In Fig. 6, we report the private leaderboard performance achieved by individual teams using (i) Track 1 Metric (T_1) and (ii) Track 2 Metric (T_2). Hereafter, we provide a short overview of the methods of the two best teams who submitted a working note (top1, top2, and top5 on the leaderboard).

upupup (Top1) [71]: The team uses a branching mechanism with gating. In total, there are three branches. They all share the first three stages of a ConvNeXt model [46], but then each branch uses different weights for the fourth stage of the ConvNeXt model. The first branch is used for the classification of all the species and is also responsible for the computation of the gating parameter. The second branch focuses on venomous snakes, while the third one focuses on harmless species. The gating parameter is used to decide which of these branches will be used. This approach is used only in the training phase and is omitted for the inference. However, the authors show that this training setup helps the model perform overall well. A combination of Seesaw loss and CE loss is used for optimization.

jack-etheredge (Top2) [17]: The team uses a CAFormer [75] model in the final solution. They introduce a new venom loss, which considers the different penalties for misclassification. A cost matrix between the predicted class and the misclassification penalty is used to reweight the softmax values of the prediction. The addition of the venom loss significantly improves the performance of the tested models across all metrics, even the F1 score. The team uses an ensemble of models trained on different data splits. Contrary to open set problems, the LogitNorm [72] did not improve the recognition rate.

ZCU-KKY (Top5) [64]: The team uses a Swin-v2 Tiny [45] model for the recognition. The reasoning behind it is so that the model can be used on mobile devices for fast and practical inference. The team combines two heads - one is for the species classification, and the other one is for venomous/harmless classification. They combine the Seesaw loss with a binary cross entropy loss. Even though the results are not as good as the results of other teams, they show an improvement over the baseline model by introducing the head responsible for venomousness recognition.

The main outcomes we can derive from the achieved results are as follows:

- An introduction of a custom loss that takes the different penalties for misclassification into account always helps. It seems to be the leading factor in improving the results.
- Branching or multi-head approach to classification of venomous vs. harmless species is another important factor in achieving better results. Although the mechanism aims to optimize the competition metric, it also improves the F1 scores. This is interesting because it shows that there are recognizable visual queues for venomousness, and it is best to model them explicitly.
- The architecture of the model (CNN vs. Transformer) is not a major cause of the success. Choosing the architecture according to other factors, such as run time or memory limitations, might be possible.

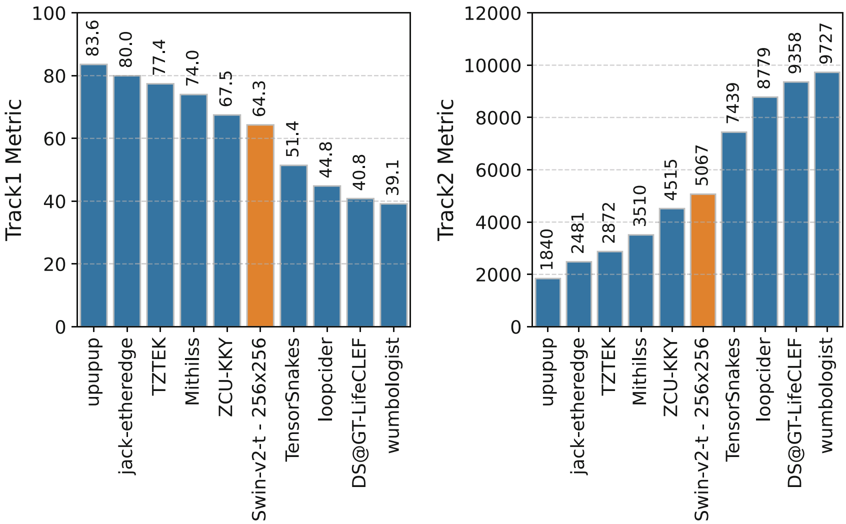


Fig. 6. Private Leaderboard – SnakeCLEF 2024 competition – All 9 teams. The orange color depicts baseline performance.

7 Conclusions and Perspectives

This new edition of LifeCLEF delivers a unique view of state-of-the-art performance on species identification and prediction problems, thanks to realistic datasets and controlled evaluation methodologies. One important conclusion is that domain shift problems remain a major problem for the emergence of new techniques such as passive acoustic sensors, HD images of plant cover, or remote sensing monitoring. The lack of annotated data for these new domains considerably hinders the progress of supervised methods, and alternative cross-domain methods are struggling to emerge. A great hope may lie in the use of unlabeled data, which will become increasingly available and whose use for domain adaptation or self-supervised learning is beginning to emerge as an effective solution (notably in BirdCLEF and GeoLifeCLEF). Another very promising prospect is multi-modal model learning, which was the key to the success of the best methods for the GeoLifeCLEF challenge and has enabled improvements in other tasks, including FungiCLEF and PlantCLEF. As far as model architectures are concerned, there is a wide disparity between the use of large-scale foundation models such as DinoV2 in PlantCLEF, SnakeCLEF, and FungiCLEF and a certain trend towards frugal architectures in GeoLifeCLEF, FungiCLEF, SnakeCLEF, and BirdCLEF. Finally, it's important to note the strength of collaborative work in the progression of the challenges. The sharing of knowledge, models, or codes, whether by the organizers or the participants themselves, has a direct impact on their subsequent developments and promotes co-construction rather than sole competition.

Acknowledgements. The research described in this paper was partly funded by the European Commission via the GUARDEN and MAMBO projects, which have received funding from the European Union's Horizon Europe research and innovation program under grant agreements 101060693 and 101060639. The opinions expressed in this work are those of the authors and are not necessarily those of the GUARDEN or MAMBO partners or the European Commission.

References

1. Affouard, A., Goeau, H., Bonnet, P., Lombardo, J.C., Joly, A.: Pl@ntnet app in the era of deep learning. In: 5th International Conference on Learning Representations (ICLR 2017), 24–26 April 2017, Toulon, France (2017)
2. Besson, M., et al.: Towards the fully automated monitoring of ecological communities. Ecol. Lett. **25**(12), 2753–2775 (2022)
3. Bolon, I., Picek, L., Durso, A.M., Alcoba, G., Chappuis, F., Ruiz de Castañeda, R.: An artificial intelligence model to identify snakes from across the world: opportunities and challenges for global health and herpetology. PLoS Negl. Trop. Dis. **16**(8), e0010647 (2022)
4. Botella, C., et al.: The GeoLifeCLEF 2023 dataset to evaluate plant species distribution models at high spatial resolution across Europe (2023)
5. Buchelt, A., et al.: Exploring artificial intelligence for applications of drones in forest ecology and management. For. Ecol. Manage. **551**, 121530 (2024)
6. Ceccaroni, L., et al.: Advancing the productivity of science with citizen science and artificial intelligence. In: Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research (2023)
7. Chai, J., Ma, Q.: Technical report for fungusclef2024 competition. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
8. Chen, Y., Peng, T., Li, W., Chen, C.S.: Combining present-only and present-absent data with pseudo-label generation for species distribution modeling. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
9. Cheng, Z., Dai, W., Sun, J.: Multi-modal feature fusion networks for geolifeclef 2024. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
10. Chiu, C., Heil, M., Kim, T., Miyaguchi, A.: Fine-grained classification for poisonous fungi identification with transfer learning. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
11. Chopard, T., Rawlings, D.: Exploring biodiversity: a multi-model approach to multi-label plant species prediction. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
12. Chulif, S., Ishrat, H.A., Chang, Y.L., Lee, S.H.: Patch-wise inference using pretrained vision transformers: neuon submission to plantclef 2024. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
13. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers (2024)
14. Das, N., Mondal, A., Chaki, J., Padhy, N., Dey, N.: Machine learning models for bird species recognition based on vocalization: a succinct review. Inf. Technol. Intell. Transp. Syst. 117–124 (2020)
15. Diao, Q., Jiang, Y., Wen, B., Sun, J., Yuan, Z.: Metaformer: a unified meta framework for fine-grained recognition. arXiv preprint arXiv:2203.02751 (2022)

16. Dyrmann, M., Mortensen, A.K., Linneberg, L., Høye, T.T., Bjerge, K.: Camera assisted roadside monitoring for invasive alien plant species using deep learning. Sensors **21**(18), 6126 (2021)
17. Etheredge, J.: Generalizable training techniques for fine-grained long-tailed image recognition: transferring methods optimized for fungiclef 2024 to snakeclef 2024. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
18. Etheredge, J.: OpenWGAN-GP for fine-grained open-set fungi classification. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
19. Foy, S., McLoughlin, S.: Utilizing dino v2 for domain adaptation in vegetation plot analysis. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
20. Gaston, K.J., O'Neill, M.A.: Automated species identification: why not? Philos. Trans. Roy. Soc. London B Biol. Sci. **359**(1444), 655–667 (2004)
21. Goëau, H., Bonnet, P., Joly, A.: Overview of PlantCLEF 2022: image-based plant identification at global scale. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (2022)
22. Goëau, H., Bonnet, P., Joly, A.: Overview of PlantCLEF 2023: image-based plant identification at global scale. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
23. Goëau, H., et al.: The imageclef 2013 plant identification task. In: CLEF task overview 2013, CLEF: Conference and Labs of the Evaluation Forum, September 2013, Valencia, Spain (2013)
24. Goëau, H., et al.: The imageclef 2011 plant images classification task. In: CLEF task overview 2011, CLEF: Conference and Labs of the Evaluation Forum, September 2011, Amsterdam, Netherlands (2011)
25. Goëau, H., et al.: Imageclef2012 plant images identification task. In: CLEF task overview 2012, CLEF: Conference and Labs of the Evaluation Forum, September 2012, Rome, Italy (2012)
26. Goëau, H., Espitalier, V., Bonnet, P., Joly, A.: Overview of PlantCLEF 2024: multi-species plant identification in vegetation plot images. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
27. Goëau, H., Lombardo, J.C., Affouard, A., Espitalier, V., Bonnet, P., Joly, A.: PlantCLEF 2024 pretrained models on the flora of the south western Europe based on a subset of Pl@ntNet collaborative images and a ViT base patch 14 dinoV2 (2024). https://doi.org/10.5281/zenodo.10848263
28. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of Wasserstein GANs. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
29. Gustineli, M., Miyaguchi, A., Stalter, I.: Transfer learning for multi-label plant species classification with self-supervised vision transformers. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
30. Joly, A., et al.: Overview of LifeCLEF 2023: evaluation of AI models for the identification and prediction of birds, plants, snakes and fungi. In: Arampatzis, A., et al. (eds.) CLEF 2023. LNCS, vol. 14163, pp. 416–439. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-42448-9_27
31. Joly, A., et al.: Overview of LifeCLEF 2018: a large-scale evaluation of species identification and recommendation algorithms in the era of AI. In: Bellot, P., et al. (eds.) CLEF 2018. LNCS, vol. 11018, pp. 247–266. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98932-7_24

32. Joly, A., et al.: Overview of LifeCLEF 2019: identification of Amazonian plants, South & North American birds, and niche prediction. In: Crestani, F., et al. (eds.) CLEF 2019. LNCS, vol. 11696, pp. 387–401. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28577-7_29
33. Joly, A., et al.: LifeCLEF 2016: multimedia life species identification challenges. In: Fuhr, N., et al. (eds.) CLEF 2016. LNCS, vol. 9822, pp. 286–310. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44564-9_26
34. Joly, A., et al.: LifeCLEF 2017 lab overview: multimedia species identification challenges. In: Jones, G.J.F., et al. (eds.) CLEF 2017. LNCS, vol. 10456, pp. 255–274. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-65813-1_24
35. Joly, A., et al.: LifeCLEF 2014: multimedia life species identification challenges. In: Kanoulas, E., et al. (eds.) CLEF 2014. LNCS, vol. 8685, pp. 229–249. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11382-1_20
36. Joly, A., et al.: LifeCLEF 2015: multimedia life species identification challenges. In: Mothe, J., et al. (eds.) CLEF 2015. LNCS, vol. 9283, pp. 462–483. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24027-5_46
37. Joly, A., et al.: Overview of LifeCLEF 2020: a system-oriented evaluation of automated species identification and species distribution prediction. In: Arampatzis, A., et al. (eds.) CLEF 2020. LNCS, vol. 12260, pp. 342–363. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58219-7_23
38. Joly, A., et al.: Overview of LifeCLEF 2022: an evaluation of machine-learning based species identification and species distribution prediction. In: Barrón-Cedeño, A., et al. (eds.) CLEF 2022. LNCS, vol. 13390, pp. 257–285. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-13643-6_19
39. Joly, A., et al.: Overview of LifeCLEF 2021: an evaluation of machine-learning based species identification and species distribution prediction. In: Candan, K.S., et al. (eds.) CLEF 2021. LNCS, vol. 12880, pp. 371–393. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-85251-1_24
40. Júnior, T.D.C., Rieder, R.: Automatic identification of insects from digital images: a survey. Comput. Electron. Agric. **178**, 105784 (2020)
41. Kahl, S., et al.: Overview of BirdCLEF 2024: acoustic identification of understudied bird species in the western ghats. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
42. Kirillov, A., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026 (2023)
43. Kong, S., Ramanan, D.: Opengan: open-set recognition via open data generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 813–822 (2021)
44. Liu, H., Tao, Z., Jiang, P., Sun, Q., Wan, M.: Plant species prediction task based on graph neural networks and cross attention methods. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
45. Liu, Z., et al.: Swin transformer v2: scaling up capacity and resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12009–12019 (2022)
46. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. arXiv preprint arXiv:2201.03545 (2022)
47. Miyaguchi, A., Aphiwetsa, P., McDuffie, M.: Tiled raster compression and embeddings for multilabel classification in geolifeclef 2024. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)

48. Miyaguchi, A., Gustineli, M., Fischer, A., Lundqvist, R.: Transfer learning with self-supervised vision transformer for snake identification. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
49. Oquab, M., et al.: Dinov2: learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
50. Picek, L., Ruiz De Castañeda, R., Durso, A.M., Sharada, P.M.: Overview of the snakeclef 2020: automatic snake species identification challenge. In: Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum (2020)
51. Picek, L., et al.: Overview of GeoLifeCLEF 2024: species presence prediction based on occurrence data and high-resolution remote sensing images. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
52. Picek, L., Durso, A.M., Bolon, I., de Castañeda, R.R.: Overview of snakeclef 2021: automatic snake species identification with country-level focus. In: Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum (2021)
53. Picek, L., Hruz, M., Durso, A.M.: Overview of SnakeCLEF 2024: revisiting snake species identification in medically important scenarios. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
54. Picek, L., Šulc, M., Chamidullin, R., Durso, A.M.: Overview of snakeclef 2023: snake identification in medically important scenarios. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
55. Picek, L., Sulc, M., Matas, J.: Overview of FungiCLEF 2024: revisiting fungi species recognition beyond 0-1 cost. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
56. Picek, L., Šulc, M., Matas, J., Heilmann-Clausen, J., Jeppesen, T.S., Lind, E.: Automatic fungi recognition: deep learning meets mycology. Sensors **22**(2), 633 (2022)
57. Picek, L., et al.: Danish fungi 2020-not just another image recognition dataset. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1525–1535 (2022)
58. Picek, L., Šulc, M., Patel, Y., Matas, J.: Plant recognition by AI: deep neural nets, transformers, and KNN in deep embeddings. Front. Plant Sci. **13**, 787527 (2022)
59. Picek, L., Durso, A.M., Hrúz, M., Bolon, I.: Overview of SnakeCLEF 2022: automated snake species identification on a global scale. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (2022)
60. Picek, L., Šulc, M., Heilmann-Clausen, J., Matas, J.: Overview of FungiCLEF 2022: fungi recognition as an open set classification problem. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (2022)
61. Picek, L., Šulc, M., Heilmann-Clausen, J., Matas, J.: Overview of FungiCLEF 2023: fungi recognition beyond 0-1 cost. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
62. Ren, H., Jiang, H., Luo, W., Meng, M., Zhang, T.: Entropy-guided open-set fine-grained fungi recognition. Aliannejadi et al. [1] pp. 2122–2136 (2023)
63. Roberts, D.R., et al.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography **40**(8), 913–929 (2017)
64. Sieber, M., Železný, T.: Do not lose to losses for snakeclef2024. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
65. Šulc, M., Matas, J.: Fine-grained recognition of plants from images. Plant Methods **13**, 1–14 (2017)
66. Sulc, M., Picek, L., Matas, J., Jeppesen, T., Heilmann-Clausen, J.: Fungi recognition: a practical use case. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2316–2324 (2020)

67. Syayfetdinov, A.: Multimodal networks for species distribution modeling. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
68. Tan, B.F., Li, Y.Y., Wang, P., Zhao, L., Wei, X.S.: Say no to the poisonous: an effective strategy for mitigating 0-1 cost in fungiclef2024. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
69. Wäldchen, J., Rzanny, M., Seeland, M., Mäder, P.: Automated plant species identification-trends and future directions. PLoS Comput. Biol. **14**(4), e1005993 (2018)
70. Wan, F., Wan, H., Zhang, Z., Gao, J., Sun, C., Wang, Y.: The application potential of unmanned aerial vehicle surveys in grassland plant diversity. Biodivers. Sci. **32**(3), 23381 (2024)
71. Wang, P., Li, Y., Tan, B.F., Zhou, Y.C., Li, Y., Wei, X.S.: Multibranch co-training to mine venomous feature representation: a solution to snakeclef2024. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
72. Wei, H., Xie, R., Cheng, H., Feng, L., An, B., Li, Y.: Mitigating neural network overconfidence with logit normalization. In: International Conference on Machine Learning, pp. 23631–23644. PMLR (2022)
73. Wolf, S., Thelen, P., Beyerer, J.: Open-set fungi classification focused on reducing risk of poisonous confusion. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
74. Yang, L., et al.: Dynamic MLP for fine-grained image classification by leveraging geographical and temporal information. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10945–10954 (2022)
75. Yu, W., et al.: Metaformer baselines for vision. IEEE Trans. Pattern Anal. Mach. Intell. (2023)
76. Zwerts, J.A., et al.: Methods for wildlife monitoring in tropical forests: comparing human observations, camera traps, and passive acoustic sensors. Conserv. Sci. Pract. **3**(12), e568 (2021)

Overview of the CLEF 2024 LongEval Lab on Longitudinal Evaluation of Model Performance

Rabab Alkhalifa[1,2], Hsuvas Borkakoty[3], Romain Deveaud[4], Alaa El-Ebshihy[5,6], Luis Espinosa-Anke[3,7], Tobias Fink[5,6], Petra Galuščáková[9], Gabriela Gonzalez-Saez[8], Lorraine Goeuriot[8], David Iommi[5], Maria Liakata[1,10,11], Harish Tayyar Madabushi[12], Pablo Medina-Alias[12], Philippe Mulhem[8], Florina Piroi[5,6(✉)], Martin Popel[13], and Arkaitz Zubiaga[1]

[1] Queen Mary University of London, London, UK
[2] Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia
[3] Cardiff University, Cardiff, UK
[4] Qwant, Paris, France
[5] Research Studios Austria, Data Science Studio, Vienna, Austria
[6] TU Wien, Vienna, Austria
florina.piroi@tuwien.ac.at
[7] AMPLYFI, Cardiff, UK
[8] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, Grenoble, France
[9] University of Stavanger, Stavanger, Norway
[10] Alan Turing Institute, London, UK
[11] University of Warwick, Coventry, UK
[12] University of Bath, Bath, UK
[13] Charles University, Prague, Czech Republic

Abstract. We describe the second edition of the LongEval CLEF 2024 shared task. This lab evaluates the temporal persistence of Information Retrieval (IR) systems and Text Classifiers. Task 1 requires IR systems to run on corpora acquired at several timestamps, and evaluates the drop in system quality (NDCG) along these timestamps. Task 2 tackles binary sentiment classification at different points in time, and evaluates the performance drop for different temporal gaps. Overall, 37 teams registered for Task 1 and 25 for Task 2. Ultimately, 14 and 4 teams participated in Task 1 and Task 2, respectively.

Keywords: Evaluation · Temporal Persistence · Temporal Generalisability · Information Retrieval · Text Classification

---

Institute of Engineering Univ. Grenoble Alpes.

L. Goeuriot et al. (Eds.): CLEF 2024, LNCS 14959, pp. 208–230, 2024.
https://doi.org/10.1007/978-3-031-71908-0_10

1 Introduction

Outside the strict scientific context, the European Artificial Intelligence Act[1], adopted by European Commission in 2024, stresses in Article 17, section (d), that providers must comply with "examination, test and validation procedures to be carried out before, during and after the development of the high-risk AI system, and the frequency with which they have to be carried out". Without focusing here on the degree of risk of Information Retrieval or Classification systems, this Act clearly states that AI systems must tackle evolution. Time is a dimension that is often overlooked when conducting Information Retrieval (IR) experiments, especially when static data sets are utilized. The advantages of such datasets are that they are easily used to evaluate and test systems. Some data sets, like CORD19, contain documents collected at different points in time, showing differences in the set of documents from one collection time to another. Recent research [15] has demonstrated that models trained on data pertaining to a particular time period struggle to keep their performance levels when applied on test data that is distant in time. On the other side, [22] showed that neural systems, especially transformers-based ones, are not always very sensitive to corpus evolution.

With the aim of tackling this challenge of making models have persistent quality over time, the objective of the LongEval lab is twofold: (i) to explore the extent to which temporal differences over time, as reflected in the evolution of evaluation datasets, results in the deterioration of the performance of information retrieval and classification systems, and (ii) to propose improved methods that mitigate performance drop by making models more robust over time.

The LongEval lab [3] took place as part of the Conference and Labs of the Evaluation Forum (CLEF) 2024, and consisted in two separate tasks: (i) Task 1, described in Sect. 2, focused on information retrieval, and (ii) Task 2, described in Sect. 3, focused on text classification for sentiment analysis. Both tasks provided labeled datasets enabling analysis and evaluation of models over data evolving in time (what we call "longitudinally evolving data").

2 Task 1 - Retrieval

The retrieval task of LongEval 2024 explores the effect of changes in datasets on retrieval of text documents. More specifically, we focus on a setup in which the datasets are evolving, as in the LongEval 2023 Retrieval Task data [3]. This means, that one dataset can be acquired from another by adding, removing (and replacing) a limited number of documents and queries. The two main scenarios considered focus on one single system or on several ones, as detailed below:

A Single System in an Evolving Setup

We explore how one selected system behaves when evaluated on several collections, which evolve along the time. The context in which this task taked place

[1] https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html.

is retrieval performances for **Web search**. When considering evolution of Web data along time, we are facing a case when the documents, the queries and also the relevance continuously evolves. We are then studying how Web search engines deal with this situation. The considered scenario is then similar to classical *ad-hoc* search, in the case of evolving data sets. The evaluation in this scenario consider both the Web search case in which the top documents are the most important elements considered, and should take into account the evolving nature of the data. Evaluation should ideally reflect the changes in the collection and especially signal substantial changes that could lead to performance drop. This would allow to re-train the search engine model then and only when it is really necessary, and enable much more efficient overall training.

As described earlier, there is no consensus about the stability of the performance of the neural networks IR systems along time, but it seems to be lower than in the case of statistical models. Moreover, the performance strongly depends on the data used for training the neural model. One objective of the task is to explore the behavior of the neural system in the evolving data scenario.

Comparison of Multiple Systems in an Evolving Setup
While in the first point, we explore a single system, comparison of this systems with multiple systems across evolving collections, should provide more information about systems stability and robustness.

2.1 Description of the Task

Compared to the LongEval 2023 Dataset [3], in 2024 we take larger lags between the training and the test sets. More precisely, the task is composed of:

– One training set, that contains Web documents, actual user's queries, and assessments, acquired at timestamp t;
– Two test sets, acquired later than t at time t' and t", composed of Web documents and user's queries.

The task datasets were created over sequential time periods, which allows doing observations at different time stamps t, and most importantly, comparing the performance across different time stamps t and t'. So, the IR task aims to assess the performance difference between t' and t" when t' occurs after t', according to teh fact that training set acquired at t, takes place few months before t'.

2.2 Dataset

As for LongEval 2023, in 2024 the data for this task were provided by the French search engine Qwant. They consist of the queries issued by the users of this search engine, cleaned Web documents, which were 1) selected to correspond to the queries, and 2) to add additional noise, and relevance judgments, which were created using a click model. The dataset is fully described in [14]. We provided training data, which included 599 train queries, with corresponding

9,785 relevance assessments and 2,049,729 Web pages. All training data were collected during January 2023. The test set corpus is composed of two subsets: Lag6 acquired in June 2023 (i.e., 6 months later than the training set), and Lag8 acquired in August 2024 (i.e. acquired 8 months later than the training set). The test dataset contains 4,321,642 documents (June: 1,790,028; August: 2,531,614) and 1,925 test queries (June: 407; August: 1,518). The datasets are accessible through the lab's webpage[2] and from the TU Wien Research Data Repository[3].

The data collected from the Qwant search engine is in French. In a way to help participants, the LongEval data set for the Retrieval task also contains automatic translations into English of both queries and documents. We mention however that the translations provided by LongEval are only applied to the first 500 characters of each sentence of the initial French documents downloaded.

The document and query overlap ratios between the collections is given by Table 1 and Table 2. We see from these tables that there is a substantial overlap between the Train and the Test collection documents and (due to the larger size of the August query set) a substantial overlap between the Train/June queries and the August queries.

Table 1. Ratio of documents shared between the LongEval 2024 train and test collections, row vs. column, i.e. 0.93 means that 93% of documents in the row collection are also included in the column collection.

	Train 2024	June (Lag6)	August (Lag8)
Train 2024	1.00	0.67	0.93
June (Lag6)	0.77	1.00	0.97
August (Lag8)	0.75	0.69	1.00

Table 2. Ratio of the queries shared between the LongEval 2024 train and test collections, rows vs. columns, i.e. 0.99 means that 99% of queries in the row collection are also included in the column collection.

	Train 2024	June (Lag6)	August (Lag8)
Train 2024	1.00	0.22	0.42
June (Lag6)	0.32	1.00	0.56
August (Lag8)	0.17	0.15	1.00

To evaluate the submissions we use one set of relevance judgments: the judgments acquired by the Qwant click model. For the evaluation, we use the NDCG measure (calculated for each dataset) at 10, as well as the drop between the Lag8 and Lag6 collection. This allows us to check to which extend the IR system face the evolution of the data. We also plan to use manual assessments, acquired through the interface described in Sect. 2.8.

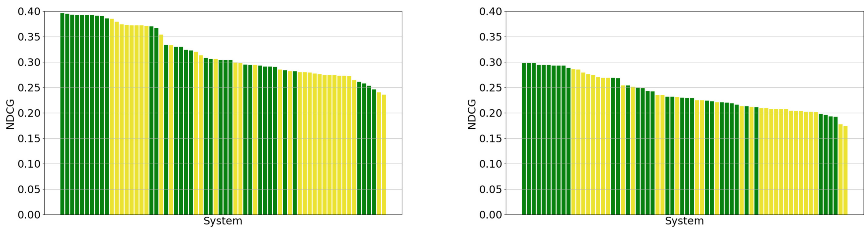
[2] https://clef-longeval.github.io/.
[3] https://doi.org/10.48436/xr350-79683.

2.3 Submissions

14 teams submitted their systems to the Retrieval task. Each team was allowed to submit up to 10 systems. Together, this a overall of 73 runs submitted (Table 9). Two teams submitted their runs on the wrong test data set, so we do not include their submission results in our further analysis.

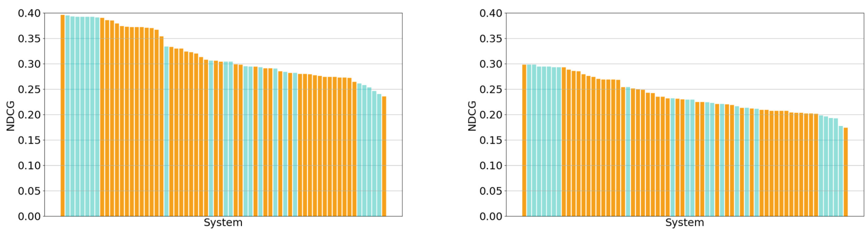
2.4 Absolute Scores

For the Retrieval task of the LongEval lab, we computed two sets of scores for each of the lags in the test collection, namely NDCG and MAP. Table 3 gives the overview of them for each run on the Lag6 and Lag8 datasets. For each run, the columns of the table indicate which language was used (English, French, or both), whether neural approaches were involved (values yes/no), and whether a single or a combination of several approaches was used (values yes/no). In addition, we show NDCG score histograms for these runs, in decreasing order, for each dataset, showing whether a run uses any neural approach (green for yes, yellow for no) in Fig. 1, and whether the run uses a combination of more than a single approach (orange for yes, cyan for no) in Fig. 2. This information was acquired from the participants through a questionnaire the participants had to fill for each submitted run. Figure 3 shows which language each made use of.



(a) Lag6 Dataset (b) Lag8 Dataset

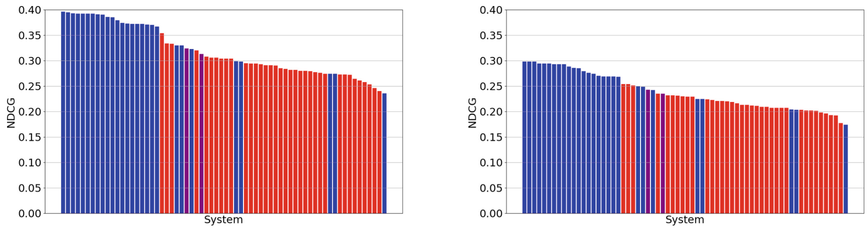
Fig. 1. Overview of the systems using a neural approach (green) vs. other (yellow). (Color figure online)



(a) Lag6 Dataset (b) Lag8 Dataset

Fig. 2. Overview of the systems which use a single approach (orange) and which use a combination of multiple approaches (cyan) (Color figure online)

From Table 3 we see that the systems which did best for the Lag6 data are also among the top for the Lag8, where the first ranked nine systems scores are comparable to each other. For instance, the best system on Lag6, according to the NDCG measure, (dam_run_4), is ranked the second best also on Lag8. Similarity, the best system on Lag8, according to the NDCG measure, (mouse_run_8), is ranked the second best also on Lag6. This finding holds for the MAP measure as well.



(a) Lag6 Dataset

(b) Lag8 Dataset

Fig. 3. Overview of the systems which use French (blue), which use English translations (red), and which use both (purple). (Color figure online)

Most of the solutions chosen by the participants to the LongEval Retrieval task apply a multi-stage retrieval approach. Often, the first stage involves a lexical-based retrieval (e.g., BM25), and query expansion methods like PL2 or BO1. Query expansion is also done by employing Large Language Models, like Mistral or Llama 3. Reranking is done either using neural-based methods or sentence based transformers. Listwise rerankers and fusing have also been used in reranking of retrieved results. Notably, the temporal aspect of the LongEval test collection has been used by some participants to include past query relevance information into query reformulation either from clicklogs or from the documents deemed relevant in the previous

Considering the Figs. 1, 2 and 3, we see that the shape of the distribution of the NDCG values are similar for the Lag6 and Lag8 datasets. However, the systems have higher performances on Lag6 than on Lag8, with maximum 0.4 value for the NDCG on the Lag6 versus 0.3 for the Lag8.

Table 3. NDCG and MAP scores for Lag6, Lag8. Results are sorted according to the NDCG scores on the Lag6.

				NDCG		MAP	
Run Id	**Neural**	**Comb.**	**Language**	**Lag6**	**Lag8**	**Lag6**	**Lag8**
dam_run_4	yes	no	French	0.396	0.294	0.249	0.171
mouse_run_8	yes	yes	French	0.395	0.298	0.248	0.174
mouse_run_10	yes	yes	French	0.393	0.298	0.246	0.175
iris_run_4	yes	yes	French	0.392	0.293	0.244	0.171
mouse_run_9	yes	yes	French	0.392	0.298	0.245	0.175
iris_run_1	yes	yes	French	0.392	0.294	0.244	0.171
iris_run_2	yes	yes	French	0.392	0.293	0.242	0.170
iris_run_3	yes	yes	French	0.391	0.293	0.243	0.171
iris_run_5	yes		French	0.390	0.294	0.240	0.171
mouse_run_7	yes	no	French	0.386	0.288	0.236	0.163
dam_run_3	no	no	French	0.385	0.285	0.235	0.162
quokkas_run_2	no	no	French	0.379	0.276	0.225	0.150
quokkas_run_1	no	no	French	0.374	0.274	0.221	0.148
lfzzo_run_7	no	no	French	0.373	0.269	0.221	0.145
lfzzo_run_8	no	no	French	0.372	0.269	0.221	0.144
lfzzo_run_9	no	no	French	0.372	0.268	0.221	0.143
lfzzo_run_10	no	no	French	0.372	0.269	0.219	0.145
lfzzo_run_6	no	no	French	0.371	0.270	0.218	0.145
dam_run_5	yes	no	French	0.370	0.279	0.220	0.156
mouse_run_6	yes	no	French	0.367	0.286	0.215	0.162
cir_run_3	no	no	English	0.354	0.242	0.226	0.136
snu_run_1	yes	yes	English	0.334	0.251	0.197	0.142
ows_run_1	no	no	English	0.333	0.243	0.199	0.139
kalu_run_2	yes	no	French	0.330	0.254	0.192	0.143
kalu_run_3	yes	no	French	0.330	0.254	0.192	0.143
kalu_run_5	yes	no	Frencg	0.324	0.249	0.188	0.140
kalu_run_4	yes	no	French	0.323	0.250	0.186	0.140
cir_run_4	no	no	English	0.320	0.229	0.172	0.117
wonder_run_3	no	no	French,English	0.313	0.235	0.163	0.116
cir_run_2	yes	no	English	0.308	0.230	0.173	0.123
mouse_run_3	yes	yes	English	0.306	0.235	0.171	0.126
ows_run_2	no	no	English	0.306	0.229	0.197	0.140
dam_run_2	yes	no	English	0.304	0.231	0.169	0.121
mouse_run_4	yes	yes	English	0.304	0.232	0.167	0.124
mouse_run_5	yes	yes	English	0.304	0.232	0.166	0.124
wonder_run_4	no	no	French	0.299	0.223	0.155	0.107
kalu_run_1	no	no	French	0.298	0.219	0.158	0.107
galapagos_run_4	yes	yes	English	0.295	0.220	0.189	0.131
ows_run_3	yes	yes	English	0.294	0.224	0.188	0.135
dam_run_1	no	no	English	0.294	0.221	0.156	0.112
galapagos_run_5	yes	yes	English	0.293	0.221	0.187	0.132
mouse_run_2	yes	no	English	0.291	0.225	0.152	0.115
mouse_run_1	yes	no	English	0.291	0.225	0.153	0.114
ows_run_7	yes	yes	English	0.290	0.213	0.180	0.123
cir_run_5	no	no	English	0.285	0.212	0.148	0.104
ows_run_6	yes	yes	English	0.284	0.216	0.173	0.126
cir_run_1	no	no	English	0.282	0.211	0.145	0.103
snu_run_2	yes	yes	English	0.282	0.213	0.177	0.127
lfzzo_run_4	no	no	English	0.280	0.209	0.142	0.102

(*continued*)

Table 3. (*continued*)

				NDCG		MAP	
Run Id	**Neural**	**Comb.**	**Language**	**Lag6**	**Lag8**	**Lag6**	**Lag8**
lfzzo_run_2	no	no	English	0.280	0.207	0.142	0.099
wonder_run_2	no	no	English	0.279	0.207	0.137	0.099
lfzzo_run_3	no	no	English	0.277	0.209	0.139	0.102
lfzzo_run_1	no	no	English	0.276	0.207	0.140	0.100
lfzzo_run_5	no	no	English	0.274	0.207	0.137	0.101
seekx_run_1	no	no	French	0.274	0.201	0.145	0.095
seekx_run_2	no	no	French	0.274	0.202	0.144	0.096
seekx_run_4	no	no	English	0.273	0.202	0.139	0.098
wonder_run_5	no	no	English	0.273	0.203	0.137	0.098
wonder_run_1	no	no	English	0.272	0.203	0.136	0.098
seekx_run_5	no	no	English	0.264	0.193	0.133	0.091
galapagos_run_2	yes	yes	English	0.261	0.198	0.162	0.115
galapagos_run_1	yes	yes	English	0.258	0.196	0.157	0.111
galapagos_run_3	yes	yes	English	0.253	0.192	0.151	0.107
ows_run_4	yes	yes	English	0.246	0.204	0.128	0.114
ows_run_5	no	yes	English	0.240	0.177	0.124	0.085
seekx_run_3	no	no	French	0.236	0.174	0.120	0.079
AVERAGE				**0.318**	**0.238**	**0.183**	**0.129**

2.5 Changes in the Scores

The main part of the retrieval task is to study the changes in the performance scores between the collections. The collections were created using the same approach and procedure have a relatively high overlap in terms of both queries and documents (see Tables 1 and 2), we thus provide the Relative NDCG Drop (RND) values of systems between the collections Lag8 and Lag6. RnD(r) for a system r, is defined as as:

$$RND(r) = \frac{\text{NDCG}_{Lag6}(r) - \text{NDCG}_{Lag8}(r)}{\text{NDCG}_{Lag6}(r)}$$

With such definition, small RND values man more robust systems against changes, and large RND values mean that the systems are not able to generalize well between lag6 and lag8. What we see in Table 4 is that the systems which are more robust to the evolution of the test collections (low values on RND) are not the best ones: for instance, ows_run_4 is the more robust system but the third worse one in Table 3. The best systems in term of NDCG values in lag6, dam_run_4 and $mouse_run$_8, have an RND of 0.245, which means that they quite robust, but much less than the most robut ones. This shows that the very best systems do cope with some extend to the evolution of the corpus, but that their is room for improving best systems against robustness. We also see that the worse robust system against changes, cir_run_3, is a system that does not rely on neural IR models: such finding shows that neural models are also likely to be more robust against changes than non-neural ones.

Table 4. Changes in the NDCG scores (RND). Lines are ordered by descending RND values. Due to the space, only the most robust run per team in terms of RND is shown.

	NDCG		RND
System	**Lag6**	**Lag8**	
ows_run_4 [1]	0.246	0.204	0.169
mouse_run_6 [9]	0.367	0.286	0.220
kalu_run_4 [19]	0.323	0.250	0.224
galapagos_run_1 [17]	0.258	0.196	0.239
dam_run_2 [8]	0.304	0.231	0.241
lfzzo_run_3	0.277	0.209	0.243
snu_run_2 [24]	0.282	0.213	0.245
wonder_run_3	0.313	0.235	0.247
iris_run_5 [13]	0.390	0.294	0.248
cir_run_2 [18]	0.308	0.230	0.252
seekx_run_4	0.273	0.202	0.260
quokkas_run_1	0.374	0.274	0.268

2.6 Run Rankings

Another point of view studied is how the submitted runs compare to each other, either in terms of the absolute NDCG scores achieved on the collections, or in terms of NDCG changes between the collections. We also calculated the Pearson correlation between the runs (now shown here), with high correlation in terms of NDCG scores, 0.99, and similarly high, 0.98, with respect to ranking order. This corresponds to the relatively high overlaps of the documents and also the queries between Lag6 and Lag8 collections (Table 1 and Table 2). This observation does not hold for the correlation between the ranking according to the NDCG score achieved and the ranking of the performance change, which is relatively low. The Pearson correlation is 0.07 for the Lag6 dataset and −0.05 on the Lag8 dataset.

Last, we calculated a combination of both rankings (ranking in terms of absolute values and ranking in terms of change). For this, we first calculated a Borda count of the ranking in terms of absolute values and Borda count of the ranking in terms of relative change and then we simply summed these two Borda counts: this result is displayed in the last column in the Table 5. We see that in terms of this measure the top performing systems (on Lag6 and Lag8 datasets) are ranked higher, although they have lower rank in terms of the rank of the NDCG change.

2.7 Queries Overview

We further investigate performance on the provided queries. Due to the space reason, we only investigate a selected subset of queries from each collection. We

Table 5. Ranking of the submitted systems by NDCG scores (columns 2–3), changes in NDCG scores between Lag6 and Lag8 dataset (column 4). Column 4 shows the sum of the Borda count applied to ranking on Lag6 and Lag8 datasets and Borda count of ranking change between Lag8 and Lag6 dataset. The darker color means better performance. Here we show the ranking of the selected runs in Table 4.

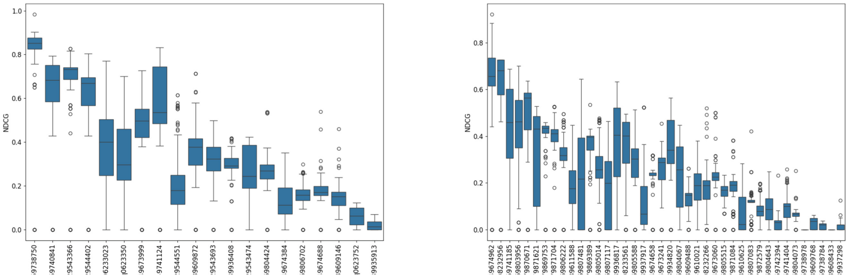
System	NDCG Lag6	NDCG Lag8	RND	Borda
iris_run_5	9	6	26	160
quokkas_run_1	13	15	57	116
mouse_run_6	20	11	2	168
kalu_run_4	27	24	3	147
wonder_run_3	29	29	25	118
cir_run_2	30	33	37	101
dam_run_2	33	32	17	119
snu_run_2	48	46	21	86
lfzzo_run_3	52	50	20	79
seekx_run_4	57	58	48	38
galapagos_run_1	62	62	13	64
ows_run_4	64	55	1	81

used a pooling strategy to select these queries to be used for the manual assessment process (described in Sect. 2.8). We first selected the top five performing runs on the average NDCG performance on both collections. We then calculated the performance of these runs per queries for each collection (i.e. Lag6 and Lag8) and sorted the queries based on their NDCG performance for the five runs. Then, we divided the query set in each collection to four sets and randomly selected from each set: five and 10 queries from Lag 6 and Lag8, respectively. We selected in total 20 queries from Lag6 collection and 40 Lag8 collection. We selected more queries from Lag8 collection since, as shown in Table 2, the number of Lag8 collection is higher than Lag6 collection.

Overview of the scores achieved for the selected queries in each collection is displayed in Fig. 4. The figure shows minimum performance (by any submitted run), 25%, quantile, 75% quantile and the maximum achieved NDCG score. Due to a relatively large number of runs, the range of the scores achieved is typically quite large and for some of the queries it even ranges between 0 and 0.8. It can be also noticed that the variation (corresponding to the size of the boxplot) of the query performance for the Lag8 collection is higher than Lag6 collection.

Some of the worst performing queries are very general ("birdsong", "taxes", and "used car" for instance) and can thus be expected to be ambiguous. This is in contrast with the top performing queries (e.g. "camping concarneau", "Prune rabbit", and "point bordeaux vision") which refer to more specific information need. Some other top performing queries have high variation in the results, e.g.

the query “origami bird” for which it is not specified if the user focuses about about"origami bird" or looks for tutorials to make them.



(a) Lag6 Dataset (b) Lag8 Dataset

Fig. 4. Selected queries performance from Lag6 and Lag8 datasets.

2.8 Manual Relevance Judgments Acquisition

The evaluation results of LongEval IR task presented above rely on automatic assessments generated from click models [14]. In addition to these click-based relevance assessments, we have set up an annotation tool to acquire further relevance assessments by humans. For that, we used the open source annotation tool, Doctag [16], on a sample of the queries selected in Sect. 2.7 (60 queries in total). Doctag provides a customizable and portable platform specifically designed for Information Retrieval (IR) evaluation. To perform manual relevance judgments using Doctag, annotators utilize its web-based interface. They access the tool and interact with its annotation functionalities, including the assignment of labels to indicate document relevance to specific queries. Annotators view the documents and associate appropriate relevance labels (Fig. 5). We set up dedicated online servers where Doctag is deployed, through their use we have acquired over 25K manual assessments which we intend to use in further evaluations.

2.9 Discussion and Conclusion

This task was the second attempt to collectively investigate the impact of the evolution of the data on search system’s performances. Having 14 participating teams submitting runs confirmed that this topic was of interest to the community.

The dataset released for this task consisted in a sequence of test collections corresponding to different times. The collections were composed of documents and queries coming from Qwant, and relevance judgment coming from a click model and manual assessment. While the manual assessment is ongoing at the time of the paper’s publication, performances of participants’ submitted runs were measured using the click logs.

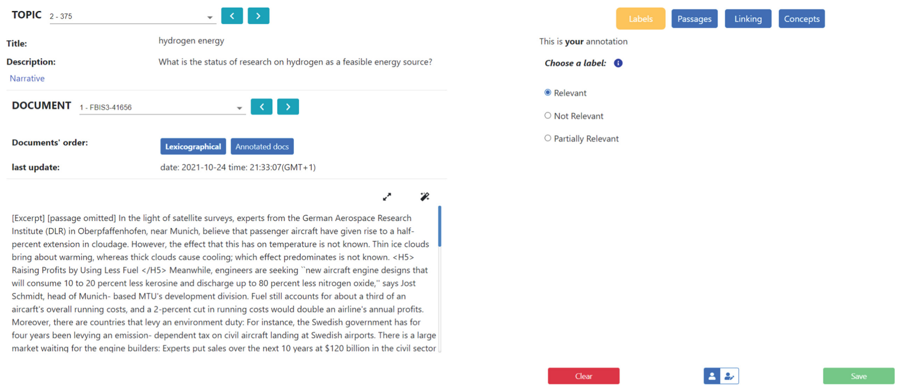


Fig. 5. Screenshot from Doctag main page. Labels annotation is done associating to each document one label that expresses the relevance of that document for that topic.

Most of submitted runs rely on multi-stage retrieval approaches. In addition to the usage of Large Language Models in Query expansion. The effect of the translation of the documents and queries provided by the lab has a clear impact: the best results were obtained on the original French data.

Since each subset had substantial overlaps, the correlations between systems rankings was pretty high. As for the robustness of the systems towards dataset changes, we observed that the systems that are the more robust to the evolution of test collection were not the best performing ones.

Further evaluations will be carried out in the near future with the manual assessment of the pooled sets. A thorough analysis of the results will be necessary to study the impact of queries on the results (their nature, topic, difficulty, etc.). Further analysis work will be necessary to fully establish the robustness of the systems and the specific impact of dataset evolution on the performances.

3 Task 2 Classification

Stance detection, an essential task in natural language processing (NLP), involves identifying an author's position or attitude towards a particular topic or statement. This task goes beyond simple sentiment analysis by requiring models to discern not just positive or negative sentiments but also the specific stance (supporting/believer, opposing/denier, or neutral) towards a given target [20, 23].

Comprehending the evolution of social media stances over time poses a significant challenge, a topic that has gained recent interest in the AI and NLP communities but remains relatively unexplored. The performance of social media stance classifiers is intricately linked to temporal shifts in language and evolving societal attitudes toward the subject matter [7].

In LongEval 2024, social media stance detection, a multi-label English classification task, takes center stage, surpassing the complexity of the binary sentiment task in LongEval 2023 [2]. Our primary goal is to assess the persistence of stance detection models in the dynamic landscape of social media posts.

The evolving nature of language and social opinions adds an additional layer of complexity to the challenges faced by text classifiers. Language undergoes continuous changes, reflecting shifts in societal norms and opinions and the emergence of novel concepts and words. For instance, consider the evolution of public opinion on climate change over the past two decades:

- **Sentence from 2000:** "Global warming is a theory that needs more proof; it's not urgent."
- **Sentence from 2010:** "Evidence for climate change is mounting, and we need to start taking action."
- **Sentence from 2020:** "Climate change is an undeniable crisis that requires immediate global action."

The context over two decades in the above example shows that language and urgency surrounding climate change have evolved from skepticism to an accepted crisis. Models not updated with recent discussions and policy changes might fail to accurately capture the critical tone and terminology used in current dialogues about the environment. Similarly, the rapid emergence of new vocabulary, as witnessed with terms like COVID-19 [6], highlights the dynamic nature of language, presenting unique challenges for text classifiers.

3.1 Description of the Task

To assess the extent of the performance drop of models over shorter and longer temporal gaps, we provided a comprehensive training dataset along with five testing sets. These testing sets include two practice sets and three development sets. The shared competition aimed to stimulate the development of classifiers that can effectively handle temporal variations and maintain performance persistence over different time distances. Participants were expected to submit solutions for two sub-tasks, showcasing their ability to address the challenges of temporal variations in performance. The shared task was in turn divided into two sub-tasks:

Sub-Task 1: Short-Term Persistence: In this sub-task, participants were tasked with developing models that demonstrated performance persistence over short periods. Specifically, the models needed to maintain their performance over a temporal gap between the within datasets and the short-term datasets. This involved comparing the performance from the **within-practice** data (January 2010 to December 2010) to the **short-practice** data (January 2014 to December 2014), a time gap of 4 years, and from the **within-dev** data (January 2011 to December 2011) to the **short-dev** data (January 2015 to December 2015), a time gap of 4 years

Sub-Task 2: Long-Term Persistence: This sub-task required participants to develop models that maintained performance persistence over a longer period of time. The classifiers were expected to mitigate performance drops over a temporal gap between the within time datasets and the long-term datasets. This involved comparing the performance from the **within-dev** data (January 2011 to December 2011) to the **long-dev** data (January 2018 to September 2019), a time gap of approximately 7 to 8 years.

In addition to the main sub-tasks, participants were also asked to work on models that maintained performance within the same temporal year of the training set, with the **practice-within** data covering January 2010 to December 2010 and the **within-dev** data covering January 2011 to December 2011, with no gap between them and the training set (time gap 0).

3.2 Dataset

In this section, we present the process of constructing our final annotated corpus for the task. The large-scale Climate Change Twitter dataset was originally described in [11], Our primary focus will be on climate change stance, time of the post (created at), and the textual content of the tweets, which we will refer to as the **CC-SD** dataset. This **CC-SD** is large-scale, covering a span of 13 years and containing a diverse set of more than 15 million tweets from various years. Using the BERT model to annotated tweets, the **CC-SD** stance labels fall into three categories: those that express support for the belief in man-made climate change (believer), those that dispute it (denier), and those that remain neutral on the topic.

The total sum of the categorized tweets over the entire time span are as follows: 11,292,424 tweets as believers, 1,191,386 as deniers, and 3,305,601 as neutral, distributed across the timeline. The annotation is performed using transfer learning with BERT as distant supervision based on another sentiment climate change dataset[4] and, thus, can be easily manually annotated to improve its precision using human in the loop.

Data Sampling. The dataset is first downsampled to ensure an equal number of instances for each stance (neutral, denier, believer) within a specified date range, using the minimum stance count across all selected months and years to avoid bias. This involves randomly sampling the same number of rows for each stance, year, and month combination, ensuring balanced representation. The downsampled data is then shuffled and split into training, development, and practice sets, including short- and long-term coverage, with any intersecting IDs between these sets being removed to maintain data integrity and prevent data leakage. Finally, a summary of the downsampled data is generated, detailing the number of rows, date and time of sampling, and statistics per year and month.

Test Set Annotation. We annotate our test data using Prolific[5], which is a high quality data collection and annotation platform. The forms that contain data to

[4] https://www.kaggle.com/datasets/edqian/twitter-climate-change-sentiment-dataset.
[5] https://www.prolific.com/.

annotate are created using Qualtrics[6]. We run the annotation in several batches, and provide the annotation guideline stating the task details and guidelines for the participants to follow. We add several filters, automatic and manual to select the optimal demographic and qualified annotators. Additionally, a manual annotation is also enforced which contains 5 tweets from the training set, which the organisers first annotate and then using the majority annotation is released as qualification task. The participant have to correctly answer 4 out of 5 questions to access the actual annotation task. We also provide fields in our form for every annotator to give their feedback and to point out if any tweet is inappropriate or contains explicit content in it. We collect responses from 5 annotators for each tweet, and select the majority annotation from the five annotation. In some cases, we find equal agreement among the annotators, and for those cases, we run an extra round of annotation to finalise the agreement. Finally after cleanup and majority annotation finding process, we manually check the data and divide into their respective splits.

The resulting distribution of data is shown in Table 6. table Dataset statistics summary of training, practice and testing sets.

Table 6. Dataset statistics summary of training, practice and testing sets.

Dataset	Time Period	Size
train	January 2009 to December 2011	35739
within-practice	January 2010 to December 2010	450
short-practice	January 2014 to December 2014	450
dev-within	January 2011 to December 2011	1074
dev-short	January 2015 to December 2015	1074
dev-long	January 2018 to September 2019	1074

In the Practice phase, participants undertake Pre-Evaluation tasks with datasets from 2010 and 2014, sampled from CC-SD, allowing them to practice within a recent time frame and over a short duration. These datasets are manually verified. Additionally, human-annotated "within time" and "short time" practice sets are provided, also sampled from CC-SD, to refine model development before formal evaluation.

Subsequently, the Evaluation phase assesses models using datasets from 2011, 2015, and the longer period of 2018–2019, all sampled from CC-SD. These datasets undergo manual verification and encompass within-timeframe assessments, short-term predictions, and long-term predictions, offering a holistic evaluation of model performance across various temporal contexts. By incorporating datasets covering different years, the evaluation ensures thorough testing and understanding of models' temporal persistence and performance.

[6] https://www.qualtrics.com/.

3.3 Evaluation

Evaluation metrics for this edition of the task remain consistent with the previous version [3,4]. All submissions were assessed using two key metrics: the **macro-averaged F1-score** on the corresponding sub-task's development set and the **Relative Performance Drop (RPD)**, calculated by comparing performance on "within time" data against results from short- or long-term distant development sets. Submissions for each sub-task were ranked primarily based on the macro-averaged F1-score. Additionally, a unified score, **the weighted-F1**, was computed between the two sub-tasks, encouraging participants to contribute to both for accurate placement on a collective leaderboard and a deeper analysis of their system's performance in various settings.

Participants were expected to design an experimental architecture to enhance a text classifier's temporal performance. In such, the performance of the submissions was evaluated in two ways:

1. **Macro-averaged F1-score:** This metric measured the overall F1-score on the testing set for the sentiment classification sub-task. The F1-score combines precision and recall to provide a balanced measure of model performance. A higher F1-score indicated better performance in terms of both positive and negative sentiment classification.

$$F_{\text{macro}} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{1}$$

2. **Relative Performance Drop (RPD):** This metric quantified the difference in performance between the "within-period" data and the short- or long-term distant testing sets. RPD was computed as the difference in performance scores between two sets. A negative RPD value indicated a drop in performance compared to the "within-period" data, while a positive value suggested an improvement.

$$RPD = \frac{f_{\text{score}_{t_j}} - f_{\text{score}_{t_0}}}{f_{\text{score}_{t_0}}} \tag{2}$$

Where t_0 represents performance when the time gap is 0, and t_j represents performance when the time gap is short or long, as introduced in previous work [5].

The submissions were ranked primarily based on the macro-averaged F1-score, emphasizing the overall performance of the stance detection model on the testing sets. The higher the macro-averaged F1-score, the higher the ranking of the submission.

3.4 Models

In our study, we evaluated several baseline classifiers to assess their performance and temporal persistence when exposed to evolving data. The models we focused on include **bert-base-uncased**, **roberta-base**, and their respective variations with additional continual incremental pretraining from the climate change corpus.

To address the challenges posed by evolving data, we implemented continual incremental pretraining for both **bert-base-uncased** and **roberta-base** models. These variations, referred to as *++MLM 2019*, were further pretrained on a climate change corpus that covers data from the initial training year up to 2019 using masked language modeling. This approach aimed to incorporate recent linguistic trends and contextual information, enhancing the models' ability to adapt to new and evolving data.

The dataset is segmented by years, starting from 2006 to various end years (2011, 2013, 2015, 2017, 2019). For each end year, data from all preceding years up to that point is aggregated and preprocessed. Preprocessing includes filling missing values with the most frequent value in each column, removing rows with missing values in the 'text' or 'stance' columns, and eliminating duplicate entries. Text data is normalized to lowercase, and entries with fewer than six words are excluded. Post-processing, the data is merged into a single dataset for each end year, resulting in five datasets representing different temporal spans. These datasets are subsequently balanced by downsampling to ensure uniform representation for incremental training.

Using a masked language modeling strategy, the textual data without its label is fed into the models incrementally in their chronological order, starting with the 2011 sample and ending with the 2019 sample. This approach ensures a balanced and clean dataset, facilitating robust analysis and model training. Each model was incrementally tested to evaluate its persistence over time, and the best performance was reported in the results section.

– **bert-base-uncased** (Bidirectional Encoder Representations from Transformers) [10] is a foundational model in NLP that introduced the concept of bidirectional training of transformers for language modeling. The bert-base-uncased model is a version of BERT that ignores case sensitivity, which helps in learning case-independent features. It also consists of 12 transformer layers, 768 hidden units, and 12 attention heads. BERT uses a static masked language modeling objective during pretraining, which involves predicting masked words in a sentence based on their context.
– **roberta-base** (Robustly optimized BERT approach) [21] is a variant of the BERT model designed to improve performance by optimizing the pretraining process. It uses dynamic masking, a larger batch size, and more data to enhance the training of transformer-based models. The roberta-base model consists of 12 transformer layers, 768 hidden units, and 12 attention heads. It is pretrained on a diverse range of data to capture rich contextual representations, making it effective for various NLP tasks.
– *++MLM 2019*: A masked language modeling strategy used to adapt a language model to new data by incrementally pretraining with an unlabeled corpus up to 2019. This method leverages recent linguistic trends and contextual updates to improve model adaptation and performance over time.

This systematic approach allowed us to evaluate and enhance the models' temporal persistence and robustness baselines, ensuring they remain effective in the face of evolving language patterns.

3.5 Results

This section presents the results obtained during both the practice and evaluation phases of task 2.

3.6 Practice Phase

In this subsection, we present the results of the practice phase of task 2. This practice dataset was provided to participants to allow them to practice and initiate their text classifiers. Since we did not get any submissions and to understand the initial performance of our practice sets, we compared several baseline classifiers. The models evaluated include **roberta-base**, **bert-base-uncased**, and their respective variations with additional continual incremental pretraining from the climate change corpus from the initial year of training up to 2019 using masked lanague modeling. The results are summarized in Table 7.

Table 7. Performance of baseline models on practice data. The columns represent: **f-Within** - performance within the same time period, **f-Short** - performance over short temporal gaps, **f-Avg** - average performance across all temporal gaps, and **RPD** - relative performance drop when applied to temporally distant data.

Model	**f-Within**	**f-Short**	**f-Avg**	**RPD**
roberta-base	0.586	0.523	0.555	-10.80%
++MLM 2019	0.612	0.525	0.569	-14.36%
bert-base-uncased	0.577	0.536	0.557	-7.19%
++MLM 2019	0.586	0.542	0.564	-7.59%

As it can be seen from Table 7, the results indicate that the *++MLM 2019* variations of both **roberta-base** and **bert-base-uncased** demonstrate improved f-Within and f-Avg scores compared to their original counterparts. This suggests that additional continual pretraining based on recent data, incrementally over time, contributes to better performance persistence. Notably, **bert-base-uncased** *++MLM 2019* achieved the lowest RPD, highlighting its resilience to temporal changes.

3.7 Evaluation Phase

In this subsection, we present the results of the evaluation phase of task 2. Using the development dataset provided to participants, we evaluated the final performance of the text classifier models. To understand the performance of our development sets, we compared several baseline classifiers due to the lack of submissions. The models evaluated include **roberta-base**, **bert-base-uncased**, and their respective variations with additional continual incremental pretraining from the climate change corpus up to 2019 using masked language modeling. The results are summarized in Table 8.

Table 8. Performance of baseline models on development sets. The columns represent: **f-Within** - performance within the same time period, **f-Short** - performance over short temporal gaps, **f-Long** - performance over long temporal gaps, **f-Avg** - average performance across all temporal gaps, **RPD-Short** - relative performance drop over short temporal gaps, **RPD-Long** - relative performance drop over long temporal gaps, and **RPD-Avg** - average relative performance drop.

Model	f-Within	f-Short	f-Long	f-Avg	RPD-Short	RPD-Long	RPD-Avg
roberta-base	0.626	0.558	0.529	0.571	-10.81%	-15.46%	-26.26%
++MLM 2019	0.623	0.594	0.552	0.590	-4.74%	-11.46%	-16.20%
bert-base-uncased	0.614	0.569	0.536	0.573	-7.26%	-12.64%	-19.89%
++MLM 2019	0.600	0.571	0.540	0.570	-4.94%	-10.01%	-14.94%

As shown in Table 8, the *++MLM 2019* variations of both **roberta-base** and **bert-base-uncased** models exhibit notable improvements in the **f-Short** and **f-Long** scores, as well as reduced **RPD** values compared to their standard counterparts. The *++MLM 2019* variation of **roberta-base** achieved an f-Avg score of (0.590), an improvement over the original model's score of (0.571). It also showed a significantly lower RPD-Short of (-4.74%) and RPD-Long of (-11.46%), indicating better resilience to temporal changes over both short and long gaps. Similarly, the *++MLM 2019* variation of **bert-base-uncased** achieved an **f-Avg** score of (0.570), slightly lower than the original model's 0.573. However, it exhibited a lower **RPD-Long** of (-10.01%) and **RPD-Avg** of (-14.94%), demonstrating improved performance persistence over time.

These results reinforce the value of continual incremental pretraining with recent data to maintain and improve model performance in dynamic environments. The *++MLM 2019* variations consistently showed enhanced performance metrics and reduced performance degradation over time, validating the effectiveness of this approach in enhancing temporal persistence.

3.8 Discussion and Conclusion

This section discusses the results of our study on temporally adaptive classification methods, highlighting the significance of incorporating temporal information into text classification models to mitigate performance drops over time and the use of an outdated language model. These results reveal that classifiers trained on older data exhibit significant performance drops when applied to newer data. This is evident from the relative performance drops (RPD) reported, where the *++MLM 2019* variations showed a marked improvement in mitigating this drop.

Previous work by Alkhalifa et al. [5] introduced the *Incremental Temporal Alignment (ITA)* method as a superior approach for enhancing temporal persistence of static word embedding. This method aligns closely with the continual incremental pretraining approach evaluated in our results, where *++MLM 2019* variations of both **roberta-base** and **bert-base-uncased** demonstrated improved **f-Within**, **f-Avg** scores, and lower **RPD** values. The *ITA* method's emphasis on leveraging incremental updates to word embeddings aligns with the improvements seen in the *++MLM 2019* models, showcasing their resilience to evolving data and enhancing their persistence as text classifiers as context updated overtime.

The results reinforce several best practices for designing temporally robust and persistent text classifiers. Methods relying on incremental updates generally outperform static embeddings, as corroborated by the superior performance of the *++MLM 2019* models. Additionally, it is crucial to select robust baseline models and incrementally update them to accommodate evolving language patterns over time.

The practical implications of our findings are significant for real-world NLP applications. In dynamic environments such as stance posts on social media, language evolves rapidly, making temporal adaptation through an incremental pretraining approach substantially enhance the longevity and persistence of text classifiers. These results provide empirical evidence supporting the implementation of temporally adaptive classification methods in real-world scenarios.

Acknowledgement. This work is supported by the ANR Kodicare bi-lateral project, grant ANR-19-CE23-0029 of the French Agence Nationale de la Recherche, and by the Austrian Science Fund (FWF, grant I4471-N). This work is also supported by a UKRI/EPSRC Turing AI Fellowship to Maria Liakata (grant no. EP/V030302/1). This work has been using services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062) and has been also supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ.

A Runs Submitted to the IR Task

Table 9. The original name of the submitted runs for the IR task are shown in the second column while the Runs Ids used assigned to the systems and used in the paper are shown in the first column.

Run Id	Submitted System
abyss_run_1	ABYSS_BM25-French-Stop50_40FR_10EN-SnowStem-Dict-Fuzzy-Phrase-Start-Synonyms-RR
abyss_run_2	ABYSS_BM25-French-Stop50_40FR_10EN-SnowStem-Fuzzy-Phrase-Start
abyss_run_3	ABYSS_BM25-French-Stop50_40FR_10EN-SnowStem-Fuzzy-Phrase-Start-RR
cir_run_1	CIR_BM25
cir_run_2	CIR_BM25+monoT5
cir_run_3	CIR_BM25+qrel_boost
cir_run_4	CIR_BM25+RF
cir_run_5	CIR_BM25+time_boost
galapagos_run_1	galapagos-tortoise-bm25-bo1-pl2-monot5-kmax-avg-k-4
galapagos_run_2	galapagos-tortoise-bm25-bo1-pl2-monot5-max
galapagos_run_3	galapagos-tortoise-bm25-bo1-pl2-monot5-mean
galapagos_run_4	galapagos-tortoise-rank-zephyr
galapagos_run_5	galapagos-tortoise-wsum
kalu_run_1	KALU_MISTRAL_FRENCH
kalu_run_2	KALU_RERANK_HARMONIC_MISTRAL_FRENCH
kalu_run_3	KALU_RERANK_HARMONIC_MISTRAL_FRENCH_SHOULD
kalu_run_4	KALU_RERANK_SIMPLE_FRENCH_LLAMA
kalu_run_5	KALU_RERANK_SIMPLE_MISTRAL_FRENCH
ows_run_1	ows_bm25_bo1_keyqueries
ows_run_2	ows_bm25_reverted_index
ows_run_3	ows_ltr_all
ows_run_4	ows_ltr_wows_all_rerank
ows_run_5	ows_ltr_wows_base_rerank
ows_run_6	ows_ltr_wows_rerank_and_keyquery
ows_run_7	ows_ltr_wows_rerank_and_reverted_index
quokkas_run_1	Quokkas_french-letter-lightstem
quokkas_run_2	Quokkas_french-standard-lightstem
dam_run_1	seupd2324-dam_EN-Stop-SnowBall-Poss-Prox(50)
dam_run_2	seupd2324-dam_EN-Stop-SnowBall-Poss-Prox(50)-Reranking(200)
dam_run_3	seupd2324-dam_FR-Stop-FrenchLight-Elision-ICU-Prox(50)
dam_run_4	seupd2324-dam_FR-Stop-FrenchLight-Elision-ICU-Prox(50)-Reranking(150)
dam_run_5	seupd2324-dam_FR-Stop-FrenchLight-Elision-ICU-Shingles-Prox(50)-Reranking(150)
iris_run_1	seupd2324-iris_FR_GFF@12_w0.162_MMARCO@1000_ADD_w5
iris_run_2	seupd2324-iris_FR_GFF@12_w0.162_MMARCO@1000_MAXMIN_ADD_w5
iris_run_3	seupd2324-iris_FR_MMARCO@1000_ADD_w5
iris_run_4	seupd2324-iris_FR_url_w1.4_GFF@12_w0.162_MMARCO@1000_ADD_w5
iris_run_5	seupd2324-iris-FR_Q2K@1_w0.16_MMARCO@1000_MAXMIN_ADD_w5
lfzzo_run_1	seupd2324-lfzzo-englishSystem1
lfzzo_run_2	seupd2324-lfzzo-englishSystem2
lfzzo_run_3	seupd2324-lfzzo-englishSystem3
lfzzo_run_4	seupd2324-lfzzo-englishSystem4
lfzzo_run_5	seupd2324-lfzzo-englishSystem5
lfzzo_run_6	seupd2324-lfzzo-frenchSystem1
lfzzo_run_7	seupd2324-lfzzo-frenchSystem2
lfzzo_run_8	seupd2324-lfzzo-frenchSystem3
lfzzo_run_9	seupd2324-lfzzo-frenchSystem4
lfzzo_run_10	seupd2324-lfzzo-frenchSystem5
mouse_run_1	seupd2324-mouse_English_Porter_Standard_NoStop_Mixtral-8x7b_NoRerank
mouse_run_2	seupd2324-mouse_English_Porter_Standard_stopwords-en_LLama3-70b_NoRerank
mouse_run_3	seupd2324-mouse_English_Porter_Standard_top125_LLama3-70b_Cohere-100-w06
mouse_run_4	seupd2324-mouse_English_Porter_Standard_top125_LLama3-70b_Pygaggle-Luyu-20-w06
mouse_run_5	seupd2324-mouse_English_Porter_Standard_top125_Mixtral-8x7b_Pygaggle-Luyu-20-w06
mouse_run_6	seupd2324-mouse_French_FrenchLight_Standard_NoStop_Mixtral-8x7b_NoRerank
mouse_run_7	seupd2324-mouse_French_FrenchLight_Standard_stopwords-fr_LLama3-70b_NoRerank
mouse_run_8	seupd2324-mouse_French_FrenchLight_Standard_top125_LLama3-70b_Cohere-100-w06
mouse_run_9	seupd2324-mouse_French_FrenchLight_Standard_top125_LLama3-70b_Pygaggle-Luyu-20-w06
mouse_run_10	seupd2324-mouse_French_FrenchLight_Standard_top125_Mixtral-8x7b_Pygaggle-Luyu-20-w06

(continued)

Table 9. (*continued*)

Run Id	Submitted System
seekx_run_1	seupd2324-seekx_LetLightFR
seekx_run_2	seupd2324-seekx_LetLightStopFR
seekx_run_3	seupd2324-seekx_LetLightStopSynFR
seekx_run_4	seupd2324-seekx_StanMinEN
seekx_run_5	seupd2324-seekx_StanMinSynEN
snu_run_1	SNU_LDI_listt5
snu_run_2	SNU_LDI_monot5
wonder_run_1	WONDER_BASELINE
wonder_run_2	WONDER_ENGLISH
wonder_run_3	WONDER_ENGLISH_FRENCH
wonder_run_4	WONDER_FRENCH
wonder_run_5	WONDER_TWOPHASE
xplore_run_1	XPLORE_French-BM25-FrenchLight-Stop
xplore_run_2	XPLORE_French-BM25-FrenchLight-Stop-SynonymMapper
xplore_run_3	XPLORE_French-BM25Default-FrenchLight-Stop
xplore_run_4	XPLORE_French-LMDirichlet-FrenchLight-Stop

References

1. Alexander, D., et al.: Team openwebsearch at clef 2024: Longeval. In: Faggioli et al. [12]
2. Alkhalifa, R., et al.: Longeval: longitudinal evaluation of model performance at CLEF 2024. In: Kamps, J., et al. (eds.) ECIR 2024. LNCS, vol. 14613, pp. 60–66. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-56072-9_8
3. Alkhalifa, R., et al.: Overview of the clef-2023 longeval lab on longitudinal evaluation of model performance. In: Arampatzis, A., et al. (eds.) CLEF 2023. LNCS, vol. 14163, pp. 440–458. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-42448-9_28
4. Alkhalifa, R., et al.: Extended overview of the clef-2023 longeval lab on longitudinal evaluation of model performance (2023). https://api.semanticscholar.org/CorpusID:259953335
5. Alkhalifa, R., Kochkina, E., Zubiaga, A.: Opinions are made to be changed: Temporally adaptive stance classification. In: Proceedings of the 2021 Workshop on Open Challenges in Online Social Networks, pp. 27–32 (2021)
6. Alkhalifa, R., Yoong, T., Kochkina, E., Zubiaga, A., Liakata, M.: QMUL-SDS at checkthat! 2020: determining COVID-19 tweet check-worthiness using an enhanced CT-BERT with numeric expressions. CoRR **abs/2008.13160** (2020). https://arxiv.org/abs/2008.13160
7. Alkhalifa, R., Zubiaga, A.: Capturing stance dynamics in social media: open challenges and research directions. Int. J. Digit. Human. **3**, 1–21 (2022)
8. Basaglia, A., Stocco, A., Popović, M., Ferro, N.: Seupd@clef: team dam on reranking using sentence embedders. In: Faggioli et al. [12]
9. Cazzador, L., Faveri, F.L.D., Franceschini, F., Pamio, L., Piron, S., Ferro, N.: Seupd@clef: team mouse on enhancing search engines effectiveness with large language models. In: Faggioli et al. [12]
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186 (2019)

11. Effrosynidis, D., Karasakalidis, A.I., Sylaios, G., Arampatzis, A.: The climate change twitter dataset. Expert Syst. Appl. **204**, 117541 (2022). https://doi.org/10.1016/j.eswa.2022.117541, https://www.sciencedirect.com/science/article/pii/S0957417422008624
12. Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.): Proceedings of Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, Aachen (2024)
13. Galli, F., Rigobello, M., Schibuola, M., Zuech, R., Ferro, N.: Seupd@clef: team iris on temporal evolution of query expansion and rank fusion techniques applied to cross-encoder re-rankers. In: Faggioli et al. [12]
14. Galuščáková, P., et al.: Longeval-retrieval: French-english dynamic test collection for continuous web search evaluation (2023)
15. Gangi Reddy, R., et al.: Synthetic target domain supervision for open retrieval QA. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1793–1797. SIGIR 2021, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3404835.3463085
16. Giachelle, F., Irrera, O., Silvello, G.: DocTAG: a customizable annotation tool for ground truth creation. In: Hagen, M., et al. (eds.) ECIR 2022 Part II. LNCS, vol. 13186, pp. 288–293. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-99739-7_35
17. Gründel, M., Weber, M., Franke, J., Reimer, J.H.: Team galápagos tortoise at longeval 2024: Neural re-ranking and rank fusion for temporal stability. In: Faggioli et al. [12]
18. Keller, J., Breuer, T., Schaer, P.: Leveraging prior relevance signals in web search. In: Faggioli et al. [12]
19. Kimia, A., Akan, A., Arwa, F., Ferro, N.: Seupd@clef: Team kalu on improving search engine performance with query expansion and re-ranking approach. In: Faggioli et al. [12]
20. Küçük, D., Can, F.: Stance detection: a survey. ACM Comput. Surv. **53**(1) (2020). https://doi.org/10.1145/3369026
21. Liu, Y., et al.: Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
22. Lovón-Melgarejo, J., Soulier, L., Pinel-Sauvagnat, K., Tamine, L.: Studying catastrophic forgetting in neural ranking models. In: Hiemstra, D., Moens, M.-F., Mothe, J., Perego, R., Potthast, M., Sebastiani, F. (eds.) ECIR 2021. LNCS, vol. 12656, pp. 375–390. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-72113-8_25
23. Mohammad, S.M., Sobhani, P., Kiritchenko, S.: Stance and sentiment in Tweets. ACM Trans. Internet Technol. **17**(3) (2017). https://doi.org/10.1145/3003433, http://alt.qcri.org/semeval2016/task6/
24. Yoon, S., Kim, J., won Hwang, S.: Analyzing the effectiveness of listwise reranking with positional invariance on temporal generalizability. In: Faggioli et al. [12]

Overview of PAN 2024: Multi-author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification Condensed Lab Overview

Abinew Ali Ayele[1], Nikolay Babakov[2], Janek Bevendorff[3(✉)], Xavier Bonet Casals[4], Berta Chulvi[5], Daryna Dementieva[6], Ashaf Elnagar[7], Dayne Freitag[8], Maik Fröbe[9], Damir Korenčić[10], Maximilian Mayerl[11], Daniil Moskovskiy[12], Animesh Mukherjee[13], Alexander Panchenko[12], Martin Potthast[14,15,16], Francisco Rangel[17], Naquee Rizwan[13], Paolo Rosso[5,18], Florian Schneider[1], Alisa Smirnova[19], Efstathios Stamatatos[20], Elisei Stakovskii[12], Benno Stein[21], Mariona Taulé[4], Dmitry Ustalov[22], Xintong Wang[1], Matti Wiegmann[21], Seid Muhie Yimam[1], and Eva Zangerle[23]

[1] Universität Hamburg, Hamburg, Germany
[2] Universidade de Santiago de Compostela, Santiago, Spain
[3] Leipzig University, Leipzig, Germany
pan@webis.de
[4] Universitat de Barcelona, Barcelona, Spain
[5] Univ. Politècnica de València, Valencia, Spain
[6] Technical University of Munich, Munich, Germany
[7] University of Sharjah, Sharjah, United Arab Emirates
[8] SRI International, Menlo Park, USA
[9] Friedrich Schiller University Jena, Jena, Germany
[10] Ruđer Bošković Institute, Zagreb, Croatia
[11] University of Applied Sciences BFI Vienna, Wien, Austria
[12] Skoltech and AIRI, Skolkovo, Russia
[13] Indian Institute of Technology Kharagpur, Kharagpur, India
[14] University of Kassel, Kassel, Germany
[15] hessian.AI, Darmstadt, Germany
[16] ScaDS.AI, Leipzig, Germany
[17] Symanto Research, Valencia, Spain
[18] ValgrAI - Valencian Graduate School and Research Network of AI, Valencia, Spain
[19] Toloka, Lucerne, Switzerland
[20] University of the Aegean, Samos, Greece
[21] Bauhaus-Universität Weimar, Weimar, Germany
[22] JetBrains, Belgrade, Serbia
[23] University of Innsbruck, Innsbruck, Austria

L. Goeuriot et al. (Eds.): CLEF 2024, LNCS 14959, pp. 231–259, 2024.
https://doi.org/10.1007/978-3-031-71908-0_11

Abstract. The goal of the PAN lab is to advance the state of the art in text forensics and stylometry through an objective evaluation of new and established methods on new benchmark datasets. IN 2024, we organized four shared tasks: (1) multi-author writing style analysis, which we continue from 2023; (2) multilingual text detoxification, a new task that aims to re-formulate text in a non-toxic way for multiple languages; (3) oppositional thinking analysis, a new task that aims to discriminate critical thinking from conspiracy narratives and identify their core actors; and (4) generative AI authorship verification, which formulates the detection of AI-generated text as an authorship problem. PAN 2024 concluded as one of our most successful editions with 74 notebook papers by 147 participating teams.

1 Introduction

PAN is a workshop series and a networking initiative for stylometry and digital text forensics. PAN hosts computational shared tasks on authorship analysis, computational ethics, and the originality of writing. Since the workshop's inception in 2007, we organized 73 shared tasks[1] and assembled 57 evaluation datasets[2] plus nine datasets contributed by the community. In 2024, we organized four tasks that concluded in 74 notebook papers by 147 participating teams.

First, the *Multi-Author Writing Style Analysis* task asks to, given a document, determine at which positions the author changes. This task was revamped for 2023 with a new dataset and structured around topical heterogeneity as an indicator of difficulty. We continued the task in 2024 with minor modifications since it attracts consistent participation of high technical quality and the problem is still relevant and offers room for improvements. A total of 15 teams submitted notebook papers to *Multi-Author Writing Style Analysis*. The task details are described in Sect. 2.

Second, the new *Multilingual Text Detoxification* task asks to, given a toxic piece of text, re-write it in a non-toxic way while saving the main content as much as possible. The task was prepared for 9 languages—English, Spanish, German, Chinese, Arabic, Hindi, Ukrainian, Russian, and Amharic—and had cross-lingual and multilingual challenges. A total of 31 teams submitted their solutions to *Multilingual Text Detoxification* resulting in 12 notebook papers. The task details are described in Sect. 3.

Third, the new *Oppositional Thinking Analysis* task asks, given an online message, to first distinguish between critical and conspiracy texts, and second, to detect the elements of the oppositional narratives. A total of 83 teams submitted their solutions to *Oppositional Thinking Analysis* resulting in 18 notebook papers. The task details are described in Sect. 4.

Fourth, the new *Generative AI Authorship Verification* task asks, given one text authored by a human and one by a machine, to pick out the human-written

[1] Find PAN's past shared tasks at http://pan.webis.de/shared-tasks.html.
[2] Find PAN's datasets at http://pan.webis.de/data.html.

one. Detecting AI-generated text is a task of high urgency and, as an authorship task, it falls deeply within PAN's expertise. We formulate AI-detection as a verification task and collaborate with the ELOQUENT Lab to generate a total of 70 different verification datasets to benchmark the PAN submissions. A total of 34 teams submitted to *Generative AI Authorship Verification*, resulting in 29 notebook papers. The task details are described in Sect. 5.

PAN is committed to reproducible research in IR and NLP, hence all participants are asked to submit their software (instead of just their predictions) through the submission software TIRA. With the recent updates to the TIRA platform [32], a majority of the submissions to PAN are publicly available as docker containers. In the following sections, we briefly outline the 2024 tasks and their results.

2 Multi-author Writing Style Analysis

The analysis of writing styles is the foundation of authorship identification tasks. The multi-author writing style analysis task, as part of PAN@CLEF, continues to develop challenges in this crucial field of research. Over the years, the task has evolved significantly: from identifying and grouping individual authors [108] to detecting whether a document has been written by a single or multiple authors [55,127,146] and identifying the actual number of authors [145], and finally, to paragraph-level style change detection [141–143].

In the PAN'24 multi-author writing style analysis task, participants were asked to identify all positions of writing style changes within a given text. Specifically, for each pair of consecutive paragraphs, the task was to compute whether there is a change in writing style between the two paragraphs. The dataset used for this task is split into three subsets of increasing difficulty: *Easy*: Each document contains a variety of topics, therefore, topic information can be used for detecting changes in writing style. *Medium:* The topics contained in a document are more homogeneous, requiring the approaches to focus more on writing style to solve the detection task. *Hard:* The paragraphs in a document are of a single topic. We control for topical diversity to ensure that, particularly in the hard dataset, topical differences cannot be used as a proxy signal for authorship and that the focus remains on stylistic cues for detecting changes in writing style.

Data Set and Evaluation

The dataset used for the multi-author writing analysis task is based on user posts on Reddit[3]. We selected posts from the following subreddits to ensure that a variety of topics is used for the creation of the datasets: *r/worldnews*, *r/politics*, *r/askhistorians*, and *r/legaladvice*. After extracting posts from these subreddits, we applied cleaning steps, such as removing quotes, whitespace, emojis, or hyperlinks. The cleaned user posts were then split into paragraphs.

[3] https://www.reddit.com/.

Table 1. Overall results for the multi-author analysis task, ranked by average F_1 performance across all three datasets. Best results are marked in bold.

Team	Easy F_1	Medium F_1	Hard F_1
fosu-stu [80]	0.987	**0.887**	0.834
nycu-nlp [68]	0.964	0.857	**0.863**
no-999 [139]	**0.991**	0.830	0.832
huangzhijian [50]	0.985	0.815	0.826
text-understanding-and-analysi [46]	**0.991**	0.815	0.818
bingezzzleep [135]	0.985	0.818	0.807
openfact [63]	0.981	0.821	0.805
chen [20]	0.968	0.822	0.807
baker [134]	0.976	0.816	0.770
gladiators [56]	0.956	0.809	0.783
khaldi-abderrahmane	0.905	0.806	0.641
karami-sh [117]	0.972	0.664	0.642
riyahsanjesh [113]	0.825	0.712	0.599
liuc0757 [72]	0.696	0.717	0.503
lxflcl66666 [66]	0.606	0.455	0.484
foshan-university-of-guangdong [73]	0.517	0.394	0.352
Baseline Predict 1	0.466	0.343	0.320
Baseline Predict 0	0.112	0.323	0.346
Baseline Random	0.414	0.506	0.495

To generate documents for the dataset, we used paragraphs from a single Reddit post to ensure minimal topical coherence between paragraphs of the generated document. Each document was composed of paragraphs written by a randomly selected number of two to four authors. For each paragraph, we extracted and computed semantic and stylistic feature vectors to characterize the paragraph. The paragraphs were then concatenated based on the similarity of their feature vectors. This mixing approach allowed us to control for topical and stylistic similarity, enabling the creation of more coherent documents and allowing us to adjust the difficulty of the multi-author writing style task. For the three datasets, we configured the similarity threshold for consecutive paragraphs to be (1) relatively large for the *easy* dataset, (2) moderate for the *medium* dataset, and (3) small for the *hard* dataset. Each of the easy, medium, and hard datasets contains 6,000 documents. We provided participants with training, test, and validation splits for all three datasets. The training sets contain 70% of the documents in each dataset, while the test and validation sets contain 15% each. The test sets were withheld for the evaluation phase of the competition.

The performance of the submitted approaches is evaluated per dataset by macro-averaged F1-score value across all documents.

Results

The task received 16 valid software submissions. The results achieved by the participants are shown in Table 1. The best average F_1 across the three datasets was achieved by the fosu-stu team. For the easy dataset, teams no-999 [139] and text-understanding-and-analysi [46] achieved the highest F_1 score (0.991), for the medium dataset, fosu-stu [80] reached an F_1 score of 0.887, and for the hard dataset, team nycu-nlp [68] achieved a F_1 of 0.863. All submissions were able to outperform the three simple baselines: a random baseline, one that predicted a style change for each pair of paragraphs, and one that predicted no style change for each pair of paragraphs. Further details on the approaches taken can be found in the overview paper [144].

3 Multilingual Text Detoxification

Text detoxification is a subtask of text style transfer where the style of text should be changed from toxic to neutral while preserving the content. As language modeling advances, there is growing concern about the potential unintended consequences of this technology. One such concern is the possibility of harmful or biased texts, which could perpetuate negative stereotypes or misinformation [64]. This has led to a growing interest in AI safety and the need for approaches to mitigating these risks [17]. This presents a major challenge for researchers and practitioners in language model safety, who need to develop effective detoxification techniques that can be applied to many languages. Previously, the first parallel corpus for such a task was released for English [75] and Russian [27] that built a foundation for the RUSSE-2022 Text Detoxification shared task.

In PAN 2024, we extend our data and challenges even to more languages. The participants were asked to develop text detoxification systems for 9 languages: English, Spanish, German, Chinese, Arabic, Hindi, Ukrainian, Russian, and Amharic. For each language, the prepared dataset was split into two parts: (i) development and (ii) test. For the train part, we did not provide any training data except for English and Russian that was publicly available from the previous work [26, 27, 75]. Thus, in the shared task, the participants were asked to do experiments in two setups:

- **Cross-lingual setup**: In the *development* phase, participants were provided 400 toxic sentences per each language. They have to experiment with various techniques for cross-lingual detoxification.
- **Multilingual setup**: Then, in the *test phase*, we released parallel dev data and asked participants to perform detoxification on 600 samples per language (3600 instances in total). At this phase, participants were able to utilize parallel training corpora to improve their approaches and perform multilingual detoxification for any subset of languages.

Table 2. The statistics of all ParaDetox datasets used in the TextDetox shared task. The human detoxified references were collected either via crowdsourcing or locally hired native speaker. For English and Russian, the previously collected train data was available during all shared task's phases. For other languages, 1 000 samples per language were divided correspondingly into development and test parts.

Language	Source of Toxic Samples	Annotation Process	Train	Dev	Test
English	[53]	Crowdsourcing+Manual	11 939	400	600
Russian	[11,115]	Crowdsourcing+Manual	8 500	400	600
Ukrainian	[16]	Crowdsourcing	—	400	600
Spanish	[96,97,124]	Crowdsourcing	—	400	600
German	[106,107,133]	Manual	—	400	600
Hindi	[82]	Manual	—	400	600
Amharic	[7,8]	Manual	—	400	600
Arabic	[40,87,89,90]	Manual	—	400	600
Chinese	[77]	Manual	—	400	600

For both phases, an automatic leaderboard was open to provide the participants scores of the adequacy and the proximity to the human references of their outputs. However, the **final** leaderboard was based on a human evaluation with crowdsourcing of subsamples from the test dataset. The human judgment gave a fair assessment of responses and prevented participants from over-tuning on automated metrics.

Data Set and Evaluation

Multilingual ParaDetox for 9 Languages. The full picture of the collected ParaDetox data for all target languages is presented in Table 2. While the methods of collecting human annotations vary across languages—some data were gathered via crowdsourcing, others by hiring local native speakers—the quality of the texts was uniformly verified by experts to ensure three key attributes as introduced in [28,75]: (i) the style of new paraphrases is genuinely non-toxic, (ii) the main content is preserved, and (iii) the new texts are fluent.

For each language for the shared task's phases:

- During the *development* phase: 400 *only* toxic parts were available for participants to perform cross-lingual experiments.
- During the *test* phase: (i) 400 ParaDetox instances were fully released; (ii) participants should provide their final solutions for 600 toxic parts of the test dataset.

For English and Russian during all phases, additional training parallel datasets were available from previous work [26,27,75]. All the data is available online for public usage.[4]

[4] https://huggingface.co/textdetox.

Automatic Evaluation. For both phases, we provided the leaderboard based on an automatic evaluation setup. We evaluate the outputs based on three parameters—style of text, content preservation, and conformity to human references—combining them into the final **J**oint score:

- **Style Transfer Accuracy (STA)** ensures that the generated text is indeed more non-toxic. It was estimated with XLM-R [22] `large` instance fine-tuned for the binary toxicity classification task for our target languages. The model determined the degree of non-toxicity in the texts.
- **Content Similarity (SIM)** is the cosine similarity between LaBSE embeddings [31] of the source texts and the generated texts.
- **Fluency (ChrF1)** is used to estimate the proximity of the detoxified texts to human references and their fluency.

Human Evaluation. We selected 100 random original toxic samples per each language from the *test* part of our dataset and performed human evaluation via Toloka crowdsourcing platform.[5] The concept of the human evaluation mirrored the approach used in the automatic evaluation. Each project type focused on assessing one of the three key qualities of detoxification; style transfer accuracy, content similarity, or fluency:

- **Style Transfer Accuracy**: we employed a pairwise comparison between the original toxic text and the generated detoxified text. Participants were tasked with determining which text was more toxic: the left text, the right text, or neither.
- **Content Similarity**: participants were shown pairs of texts (toxic phrase followed by detoxified phrase) and asked to indicate if the sense was similar, responding with "yes" or "no".
- **Fluency**: individual sentences were evaluated for intelligibility and correctness. Annotators could respond with "yes", "partially", or "no", corresponding to scores of 1, 0.5, and 0, respectively. The fluency score for a text pair was determined by comparing the detoxified text's score to the original. If the detoxified text had a higher or equal fluency score, the pair received a 1; otherwise, it received a 0.

Final Joint Score (J). For both automatic and human evaluation setups, the J score was the aggregation of the three above metrics. The metrics **STA**, **SIM** and **FL** were subsequently combined into the final **J** score used for the final ranking of approaches. Given an input toxic text x_i and its output detoxified version y_i, for a test set of n samples:

$$\mathbf{J} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{STA}(y_i) \cdot \mathbf{SIM}(x_i, y_i) \cdot \mathbf{FL}(x_i, y_i),$$

[5] https://toloka.ai.

Table 3. Results of the *human* final evaluation of the TextDetox test phase. Scores are sorted by the average **J**oint score. Scores are sorted by the average **J**oint score across all 9 languages. Baselines are highlighted with gray, Human References are highlighted with green.

Team	**Avg**	**System**
Human References	0.851	Human paraphrases from our multilingual ParaDetox
SomethingAwful	0.774	Few-shot LLaMa-3 prompting+mT0-XL
adugeen	0.741	Fine-tuned mT0-XL with ORPO [43]
VitalyProtasov	0.723	Preprocessing+mT0-large
nikita.sushko	0.712	Fine-tuned mT0-XL+postprocessing
erehulka	0.708	Few-shot LLaMa-3 prompting
bmmikheev	0.685	Few-shot LLaMa-3 prompting+GPT-3.5 post-eval.
mkrisnai	0.681	Few-shot GPT-3.5 prompting
d1n910	0.654	Few-shot Kimi.AI prompting
Yekaterina29	0.639	Fine-tuned mT5-XL
estrella	0.576	Tree of Thought GPT3.-5 prompting
gleb.shnshn	0.564	Zero-shot LLaMa-3-70b prompting
Delete	0.560	Elimination of toxic keywords
mT5	0.541	Fine-tuned mT5-XL
shredder67	0.524	Fine-tuned mT5-XL
razvor	0.516	Few-shot LLaMa-3 prompting
ZhongyuLuo	0.513	Translation+BART-detox&ruT5-detox
gangopsa	0.500	Fine-tuned T5&BART+token-level editing
Backtranslation	0.411	Translation of data to English+BART-detox
maryam.najafi	0.177	Mistral-7b with PPO
dkenco	0.119	Few-shot Cotype-7b prompting

where $\mathbf{STA}(y_i)$, $\mathbf{SIM}(x_i, y_i)$, $\mathbf{FL}(x_i, y_i) \in [0, 1]$ for *automatic* and $\in \{0, 1\}$ for *human* evaluation for each text detoxification output y_i.

We calculated all the metrics separately per each language. In the end, we calculated the **Average** score of 9 **J**oint scores per all languages that were used to compile the leaderboard.

Results

We received 20 submissions for the development phase leaderboard and 31 submissions for the test phase leaderboard; the final manually evaluated leaderboard was based on 17 submissions who confirmed their participation in the competition [34,63,78,91,93,95,99,102,105,110,123,130,149]. The final leaderboard based on human assessments is presented in Table 3.

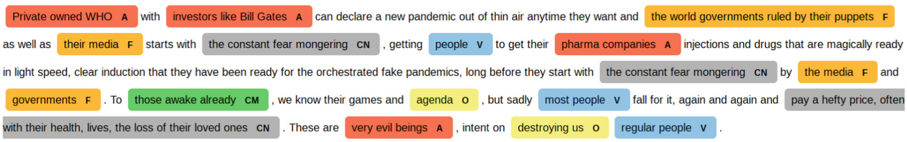


Fig. 1. A Telegram text annotated with elements of oppositional narrative.

Almost all of the participants used the current SOTA LLMs, among which are GPT-3.5 [92] and Llama-3 [3] models; to enhance the model's performance on the task of detoxification participants tested both zero-shot and few-shot prompting methods. Among smaller models, there were used mT5 [137] and mT0 [88]—these models were usually finetuned using ad hoc filtering and data augmentation techniques, for instance, as RAG and backtranslation. Additionally, region-specific LLMs were also employed: Cotype-7b [86] and Kimi.AI [2].

The majority of the participants overcame the baselines and even a couple of solutions outperformed human references. Still, for not-so-rich-resource languages such as Ukrainian, Chinese, Amharic, and Hindi human detoxified paraphrases remained the gold standard. At the same time, various experiments from participants illustrate that vanilla usage of LLMs for the detoxification task does not achieve high results. At least more advanced prompting techniques and fine-tuning on the downstream task with our provided data boosted the performance significantly achieving such interesting SOTA results.

4 Oppositional Thinking Analysis: Conspiracy Theories vs Critical Thinking Narratives

Conspiracy theories are complex narratives that attempt to explain the ultimate causes of significant events as cover plots orchestrated by secret, powerful, and malicious groups [29]. A challenging aspect of identifying conspiracy with NLP models [33,35,62,100,101,109] stems from the difficulty of distinguishing critical thinking from conspiratorial thinking in automatic content moderation. This distinction is vital because labeling a message as conspiratorial when it is only oppositional could drive those who were simply asking questions into the arms of the conspiracy communities.

At PAN 2024 we aim at analyzing oppositional thinking, and more concretely, at discriminating conspiracy from critical narratives from a *stylometry* perspective. The task will address two new challenges for the NLP research community: (1) to distinguish the conspiracy narrative from other oppositional narratives that do not express a conspiracy mentality (i.e., critical thinking); and (2) to identify in online messages the key elements of a narrative that fuels the inter-group conflict in oppositional thinking. Accordingly, we propose two sub-tasks:

- **Subtask 1** is a binary classification task differentiating between (1) critical messages that question major decisions in the public health domain, but

do not promote a conspiracist mentality; and (2) messages that view the pandemic or public health decisions as a result of a malevolent conspiracy by secret, influential groups.
- **Subtask 2** is a token-level classification task aimed at recognizing text spans corresponding to the key elements of oppositional narratives. Since conspiracy narratives are a special kind of causal explanation, we developed a span-level annotation scheme that identifies the goals, effects, agents, and the groups-in-conflict in these narratives.

For the second task, a new fine-grained annotation scheme was developed with the goal of identifying, at the text span level, how oppositional and conspiracy narratives use inter-group conflict. The annotation was performed for the described 5,000 binary-labeled messages per language. We identify the following six categories of narrative elements at the span level (see Fig. 1):

- **Agents**: the hidden power that pulls the strings of the conspiracy. In critical messages, agents are actors that design the mainstream public health policies: Government, WHO, ...;
- **Objectives**: parts of the narrative that answer the question "What is intended by the agents of the conspiracy theory or by the promoters of the action being criticized from a critical thinking perspective?";
- **Consequences**: parts of the narrative that describe the effects of the agent's actions;
- **Facilitators**: the facilitators are those who collaborate with the conspirators; in critical messages, facilitators are those who implement the measures dictated by the authorities;
- **Campaigners**: in conspiracy messages, the campaigners are the ones who uncover the conspiracy theory; in critical messages, campaigners are those who resist the enforcement of laws and health instructions; and
- **Victims**: the people who are deceived into following the conspiratorial plan or the ones who suffer due to the decisions of the authorities.

Data Set and Evaluation

For the creation of the corpus, we first manually compiled a list of 2,273 public Telegram channels in **English** and **Spanish** that contain oppositional non-mainstream views on the COVID-19 pandemic. We retrieved and filtered messages from the channels based on a set of oppositional and conspiracy keywords related to COVID-19. Then the messages were cleaned by removing duplicates, short texts, and texts with a large proportion of non-regular words (such as URLs and mentions). Finally, the messages were ranked using an index of quality based on the properties of a message and its channel. The index is composed of several criteria capturing the prevalence of COVID-19 topics and the channel's activity.

We developed an annotation schema to differentiate between the messages criticizing the mainstream views on COVID-19 and the messages evoking the existence of a conspiracy. A message was labeled "conspiracy" if any of these four

Table 4. Overall results for subtask 1 on Conspiracy theories vs Critical thinking narratives in English (EN) in terms of Matthew's correlation coefficient (MMC).

Team	EN MCC	Team	EN MCC	Team	EN MCC
IUCL	0.838	nlpln	0.784	lnr-lladrogal	0.725
AI_Fusion	0.830	RalloRico	0.777	lnr-fanny-nuria	0.725
SINAI	0.829	LasGarcias	0.775	MarcosJavi	0.719
ezio	0.821	zhengqiaozeng	0.775	lnr-cla	0.716
hinlole	0.819	ALC-UPV-JD-2	0.772	lnr-jacobant.	0.716
Zleon	0.819	LorenaEloy	0.771	MUCS	0.716
virmel	0.819	lnr-alhu	0.770	lnr-aina-julia	0.715
inaki	0.814	NACKO	0.769	LaDolceVita	0.707
yeste	0.812	paranoia-pulverizers	0.768	alopfer	0.705
auxR	0.808	DiTana	0.765	lnr-luqrud	0.705
Elias&Sergio	0.803	FredYNed	0.764	LNR-JoanPau	0.705
theateam	0.803	dannuchihaxxx	0.764	lnr-carla	0.700
trustno1	0.798	lnr-detectives	0.763	lnr-Inetum	0.698
DSVS	0.797	TargaMarhuenda	0.761	lnr-antonio	0.685
sail	0.796	Trainers	0.759	LluisJorge	0.678
ojo-bes.	0.796	thetaylorswift	0.757	anselmo-team	0.672
RD-IA-FUN	0.796	locasporlnr	0.757	lnr-pavid	0.595
Baseline BERT	0.796	lnr-adri	0.755	LNRMADME	0.546
aish_team	0.791	TokoAI	0.754	lnr-mariagb.	0.506
rfenthusiasts	0.790	ede	0.753	LNR_08	0.442
Dap_upv	0.789	lnr-verdnav	0.752	Kaprov	0.370
oppositional_opposition	0.789	lnr-dahe	0.748	lnr_cebusqui	0.048
RD-IA-FUN	0.789	epistemologos	0.748	jtommor	0.040
miqarn	0.788	lucia&ainhoa	0.747	eledu	0.459
CHEEXIST	0.787	pistacchio	0.741	david-canet	0.631
tulbure	0.787	lnr-BraulioP.	0.739	lnr-guilty	0.659
XplaiNLP	0.787	Marc_Coral	0.739	lnrANRI	0.755
TheGymNerds	0.785	Ramon&Cajal	0.728	ROCurve	0.800

criteria were met: (1) it framed COVID-19 or a related public health strategy as the result of the agency of a small and malevolent secret group; (2) it claimed that the pandemic is not real (e.g. a plandemic); (3) it accused critics of the conspiracy theory of being a part of the plot; (4) it divided society into two: those who know the truth (the conspiracy theorists) and those who remain ignorant. A message was labeled "critical" if it opposed publicly accepted understandings of events but had none of these four characteristics of the conspiratorial mindset.

Using this annotation scheme, 5,000 messages per language were annotated as "conspiracy" or "critical" thinking. For these messages, we performed anonymization by removing sensitive and identifiable information such as nicknames, user

Table 5. Overall results for subtask 1 on Conspiracy theories vs Critical thinking narratives in Spanish (ES) in terms of Matthew's correlation coefficient (MMC).

Team	ES MCC	Team	ES MCC	Team	ES MCC
SINAI	0.742	NACKO	0.646	Ramon&Cajal	0.58
auxR	0.720	ALC-UPV-JD-2	0.646	lnr-fanny-nur.	0.58
RD-IA-FUN	0.702	DSVS	0.646	lnr-antonio	0.57
Elias&Sergio	0.697	RD-IA-FUN	0.644	LluisJorge	0.56
AI_Fusion	0.687	locasporlnr	0.643	lnr-cla	0.56
zhengqiaozeng	0.687	DiTana	0.637	lnr-jacobant.	0.56
virmel	0.685	lnr-BraulioPaula	0.635	lnr-pavid	0.55
trustno1	0.684	Dap_upv	0.630	alopfer	0.55
Zleon	0.682	TheGymNerds	0.630	LNRMADME	0.54
ojo-bes	0.681	MUCS	0.629	lnr-carla	0.54
tulbure	0.672	LasGarcias	0.624	LorenaEloy	0.54
sail	0.671	lnr-dahe	0.619	CHEEXIST	0.53
nlpln	0.668	lnr-adri	0.619	lnr-guilty	0.52
Baseline BERT	0.668	hinlole	0.619	eledu	0.50
pistacchio	0.667	RalloRico	0.610	lnr-mariagb.	0.49
rfenthusiasts	0.665	lnr-aina-julia	0.61	dannuchihaxxx	0.47
XplaiNLP	0.662	lnr-verdnav	0.61	lnr-detectives	0.40
yeste	0.660	thetaylorswift	0.60	LNR_08	0.06
oppositional_opposition	0.660	lnr-alhu	0.60	jtommor	0.01
epistemologos	0.656	lnr-luqrud	0.60	lnr-Inetum	0.00
miqarn	0.656	lnr-lladrogal	0.59	Marc_Coral	0.00
theateam	0.655	ede	0.59	MarcosJavi	-0.03
ezio	0.653	Fred&Ned	0.59	lnr_cebusqui	-0.41
lucia&ainhoa	0.652	LaDolceVita	0.59	david-canet	-0.50
TargaMarhuenda	0.651	LNR-JoanPau	0.59	lnrANRI	-0.61
TokoAI	0.651	anselmo-team	0.58	ROCurve	-0.64
paranoia-pulver.	0.649				

IDs, and e-mail addresses. The average text length is 128 tokens for Spanish texts and 265 tokens for English texts that tend to elaborate more on conspiracy theories.

Each message was annotated by three linguists and the inter-annotator agreement (IAA) was calculated. Disagreements were discussed with the social psychologist who created the annotation scheme. For English messages, the IAA in terms of Krippendorf's α is 0.79 for "conspiracy" messages and 0.60 for "critical" messages, while the average observed percentage of agreement between the three annotators is 91.4%, and 80.3%, respectively. For Spanish messages, Krippendorf's α is 0.80 for "conspiracy" messages and 0.70 for "critical" messages, corresponding to the percentage agreements of 90.9% and 84.9%.

Table 6. Overall results for subtask 2 on the Text-span recognition of elements of oppositional narratives, in English (EN) and Spanish (ES), in terms of macro-averaged span-F1

Team	EN span-F1
tulbure	0.6279
Zleon	0.6089
hinlole	0.5886
oppositional_opposition	0.5866
AI_Fusion	0.5805
virmel	0.5742
miqarn	0.5739
TargaMarhuenda	0.5701
ezio	0.5694
zhengqiaozeng	0.5666
Elias&Sergio	0.5627
DSVS	0.5598
CHEEXIST	0.5524
rfenthusiasts	0.5479
ALC-UPV-JD-2	0.5377
Baseline BETO	0.5323
Dap_upv	0.5272
aish_team	0.5213
SINAI	0.4582
Trainers	0.3382
nlpln	0.3339
ROCurve	0.2996
TokoAI	0.2760
DiTana	0.2756
TheGymNerds	0.2070
epistemologos	0.1709
theateam	0.1503
LaDolceVita	0.0726
kaprov	0.0150

Team	ES span-F1
tulbure	0.6129
Zleon	0.5875
AI_Fusion	0.5777
CHEEXIST	0.5621
virmel	0.5616
miqarn	0.5603
DSVS	0.5529
TargaMarhuenda	0.5364
Elias&Sergio	0.5151
hinlole	0.4994
Baseline BETO	0.4934
Dap_upv	0.4914
zhengqiaozeng	0.4903
ALC-UPV-JD-2	0.4885
ezio	0.4869
nlpln	0.4672
rfenthusiasts	0.4666
SIANI	0.4151
TheGymNerds	0.3984
DiTana	0.3004
ROCurve	0.2649
TokoAI	0.1878
epistemologos	0.1657
LaDolceVita	0.1056
theateam	0.0994
oppositional_opposition	0.0037

For the second task, a new fine-grained annotation scheme was developed with the goal of identifying, at the text span level, how oppositional and conspiracy narratives use intergroup conflict. The annotation was performed for the described 5,000 binary-labeled messages per language.

In the process of span-level annotation, each of the 5,000 Spanish and English messages were annotated by two linguists. Currently, the annotation instructions are being discussed and improved and, to this end, we are using the Gamma (γ) measure of the IAA test [83], yielding a first average γ of 0.43. The following batch had an average gamma of 0.53, and the last one had a γ of 0.61. We deemed this a good agreement because it is close to or above the average agreement of other highly conceptual span-level schemes [24,132]. A detailed description of the dataset can be found in [60].

The official evaluation metric for subtask 1 (critical vs. conspiracy classification) is Matthew's correlation coefficient (MCC) [21], while the official metric for subtask 2 (span-level detection of narrative elements) is macro-averaged span-F1 [23].

Results

A total of 83 teams submitted their runs for subtasks 1 and 2, resulting in 18 notebook submissions [4,6,9,25,30,36,44,47,51,71,81,111,112,128,131,147,150]. In the tables above we illustrate the ranking per language. Concretely, Table 4 and Table 5 show the overall results obtained for subtask 1 on Conspiracy theories vs critical thinking narratives, in terms of Matthew's correlation coefficient; while Table 6 shows the results of subtask 2 on Text-span recognition of elements of oppositional narratives, in terms of macro-averaged span-F1.

We will analyze in detail the results and describe the models of the participants in the task overview paper [61].

5 Voight-Kampff Generative AI Authorship Verification

Authorship verification is a fundamental task in author identification. All cases of questioned authorship can be decomposed into a series of verification instances, be it in a closed-set or open-set scenario [59]. Since PAN has been continuously organizing Authorship verification tasks [12,13,118,119], we are well-equipped to tackle a timely and highly important issue: identification of machine authorship in contrast to human authorship.

Authorship identification of generative AI "in the wild" where a single document is disputed without reference is an open-set problem and the hardest formulation of the task. Although the literature suggests limited success in solving this problem given the current generation of LLMs, it is questionable whether this will remain so with improving technology. Setting aside mixed human and machine authorship, we have broken down all possible formulations of the problem with increasing levels of difficulty to get a more fundamental understanding of the task at hand and the feasibility of potential solutions. Figure 2 visualizes the cascade of all problem variants from easiest (Task 1) to most difficult (Task 7). In the easiest case, two documents with unknown authorship are given, yet we guarantee that exactly one is generated by a human $\boxed{\text{A}}$, and the other by a machine $\boxed{\text{M}}$, respectively. This constraint is relaxed in the following variants

Input / Task		**Possible Assignment Patterns**
1. $\{ \boxed{?}, \boxed{?} \}$		1. $\{ \boxed{A}, \boxed{M} \}$
2. $\{ \boxed{?}, \boxed{?} \}$		2. $\{ \boxed{A}, \boxed{M} \}, \{ \boxed{A}, \boxed{A} \}$
3. $\{ \boxed{?}, \boxed{?} \}$	$\longrightarrow$	3. $\{ \boxed{A}, \boxed{M} \}, \{ \boxed{M}, \boxed{M} \}$
4. $\{ \boxed{?}, \boxed{?} \}$		4. $\{ \boxed{A}, \boxed{M} \}, \{ \boxed{A}, \boxed{A} \}, \{ \boxed{M}, \boxed{M} \}$
5. $\{ \boxed{?}, \boxed{?} \}$		5. $\{ \boxed{A}, \boxed{M} \}, \{ \boxed{A}, \boxed{A} \}, \{ \boxed{A}, \boxed{B} \}$
6. $\{ \boxed{?}, \boxed{?} \}$		6. $\{ \boxed{A}, \boxed{M} \}, \{ \boxed{A}, \boxed{A} \}, \{ \boxed{A}, \boxed{B} \}, \{ \boxed{M}, \boxed{M} \}$
7. $\boxed{?}$		7. $\boxed{A}$, $\boxed{M}$

Fig. 2. Hierarchy of authorship verification problems from "easiest" (1) to "hardest" (7), involving LLM-generated text. Ignoring mixed human and machine authorship, the difficulty arises from the pairing constraints imposed by the possible assignment patterns. $\boxed{M}$ denotes LLM-generated text, while $\boxed{A}$ and $\boxed{B}$ denote human-authored text (same letter meaning same human author).

where, for example, both texts may also stem from a machine, $\{\boxed{M}, \boxed{M}\}$. In the hardest case, a single text is given, which could be either $\boxed{A}$ or $\boxed{M}$.

For the 2024 task on "Generative AI Authorship Verification," we follow the "easiest" formulation of the task in order to establish a feasibility baseline. The task description reads: "Given two texts, one authored by a human, one by a machine: pick out the human."

The task is organized in collaboration with the ELOQUENT Lab [54] in a builder-breaker style, in which PAN participants build systems to identify machine authorship, while ELOQUENT participants supply datasets trying to break the systems.

Data Set

In addition to the ELOQUENT-provided data, we collected 1,359 articles of major 2021 U.S. news headlines from Google News. We chose this time period specifically as it predates the release of GPT-3.5 so that we could be reasonably certain the articles were actually human-authored. We used GPT-4-Turbo to generate a bullet-point summary of each article and the summaries were then given to a selection of 13 downstream large language models to write new articles from them.

Of the original 1,359 human-authored articles, participants were given 1,087 together with their machine counterparts from 13 LLMs to calibrate their systems. The remaining 272 articles and generations from 15 LLMs were kept back for testing, resulting in 3,984 test cases, which together form the "main" portion of the test set.

To further test the robustness of the submitted systems, we generated multiple variants of the original pairs. In particular, we: (1) amended the prompt to generate German instead of English texts (this was already part of the "main"

Table 7. Overview of the 65 dataset variants provided as baseline datasets. All variants contain the same 271 human texts and (roughly) one machine generated text per LLM used. Discarding erroenous generations, this results in 3,441 pairings each for main and cross-domain variants, 600 for both unicode variants and short texts, 543 for german texts, 542 for the Kaggle prompt, 272 for both contrastive decoding (* using Llama2-13B).

	ChatCat	**Bloomz**	**Gemini Pro** *with temp.*		**Text-Bison**	**GPT**			**Llama2**		**Mistral**		**Qwen-1.5**
Variation / Obfuscation	7B	7B	0.6	0.9	002	2-OI	3.5	4	7B	70B	7B	8x7B	72B
Main	x	x	x	x	x	x	x	x	x	x	x	x	x
Unicode sub. (machine)			x	x		x	x	x	x	x			
Unicode sub. (both)			x	x		x	x	x	x	x			
Cross-domain	x	x	x	x	x	x	x	x	x	x	x	x	x.
Short text	x	x	x	x	x	x	x	x	x	x	x	x	x
German text (machine)					x			x					
Contr. decoding ($\alpha = 0.1$)									x*				
Contr. decoding ($\alpha = 0.6$)									x*				
Kaggle prompt			x					x					

test set, but not communicated to the participants); (2) replaced 15 % of the characters in (a) the machine texts and (b) both the human and machine texts with Unicode lookalike characters; (3) shuffled the test case pairs to break the topic coherence; (4) used contrastive decoding [121] instead of top-k / top-p sampling; (5) cropped texts to 35 words; and (6) used the prompt from a previous Kaggle competition on LLM detection [57] to generate more faithful paraphrases of the original articles, instead of using the stripped-down bullet point summaries.

In total, we created 65 test set variations from 13 (15) different LLMs, which are summarized in Table 7, with ELOQUENT providing another five. A more detailed description is available in the joint task overview paper [15].

Evaluation

At test time, participants were given pairs of human and LLM texts and had to calculate a score between 0 and 1, indicating which text was more likely to be human-authored. Scores less than 0.5 mean the left text is human and scores greater than 0.5 mean the right text is human. A score of exactly 0.5 could be given to signal a non-decision. We borrowed this evaluation scheme from previous installments of the PAN Authorship Verification Task.

We rank systems by their macro-average effectiveness across all $n = 70$ dataset variants (including ELOQUENT submissions) discounted by half a standard deviation (estimated from the scores with $n-1$ DoF), which penalizes unstable systems that are not robust against text obfuscations or other text variations. We use the macro average over datasets since all datasets have different numbers of examples, yet we consider them equally important as performance indicators.

Also in line with previous task installments, we compute the effectiveness for each dataset variant as the average of the established evaluation measures in authorship verification (all with comparable 0–1 scales). In particular:

- Roc-Auc: The area under the Receiver Operating Characteristic curve.
- Brier: The complement of the Brier score (mean squared loss)
- C@1: A modified accuracy score that assigns non-answers (score = 0.5) the average accuracy of the remaining cases.
- F_1: The harmonic mean of precision and recall.
- $F_{0.5u}$: A modified $F_{0.5}$ measure (precision-weighted F measure) that treats non-answers (score = 0.5) as false negatives.

Submitted Systems

In total, our task attracted 34 teams to submit systems in addition to the baseline systems we provided. Table 8 shows the best-performing system of each team that submitted notebook papers and a brief description of their approach.

Baselines

We provided implementations of six baseline systems to compare submitted systems against four state-of-the-art zero-shot LLM detection baselines and two adapted authorship verification baselines.

The zero-shot LLM detection baselines are: (1) Binoculars [42], (2) DetectLLM (both NPR and LRR scoring mode), (3) DetectGPT [85], and (4) Fast-DetectGPT [10]. All three were provided in two variants using either Falcon-7B [5] or Mistral-7B [52] to estimate text perplexities. The required text perturbations for DetectGPT and DetectLLM-NPR were generated with T5-3B [104].

The two authorship verification baselines were adapted to the LLM detection task by splitting each text in half and comparing the two halves against each other under the assumption that LLM texts are stylistically more self-similar than human texts. The baselines provided are a compression model (PPMd CBC) [41,114] and short-text authorship unmasking [14,58].

As an additional seventh baseline, we measured and compared the text lengths in characters. This baseline serves as both a quasi-random baseline and as a data sanity check.

Participant Systems. While our baseline systems reproduce established methods in either authorship verification or intrinsic, zero-shot LLM detection, the participant systems cover a broad range of approaches. The most popular approach is to use a BERT-based classifier with some modification (like PU loss or R-Drop), bagging, and/or expansion of the given training data with other

Table 8. The score is the mean of all evaluation measures across all other metrics on the main dataset corrected by half a standard deviation to correct for spread.

Team	Score	System
Tavan [125]	0.924	Ensemble: LoRA-trained LLM + Binoculars
J. Huang [46]	0.921	BERT with multiscale PU loss [126]
Lorenz [76]	0.886	SVM with TF-IDF features
M. Guo [39]	0.884	LSTM embeddings + GPT-2 PPL
Z. Lin [69]	0.851	Finetuned BERT + R-Drop
Abburi [1]	0.843	Ensemble: RoBERTa + E5 + GPT-2 Perplexity
Miralles [84]	0.806	Entropy and text features + XGBoost
Yadagiri [138]	0.806	Finetuned BERT + linguistic features
Lv [79]	0.804	Finetuned DeBERTa with Reptile meta learning
Gritsai [37]	0.796	Ensemble: LoRA-trained LLMs
Cao [18]	0.778	*Finetuned BERT*
L. Guo [38]	0.763	BERT and text features + Bi-LSTM
Binoculars 1	0.741	*Baseline Binoculars (Falcon-7B) [42]*
B. Huang [45]	0.735*	Finetuned BERT + R-Drop [67]
Valdez-Valenzuela [129]	0.727*	Graph Neural Network + BERT
Ye [140]	0.722	T5 with LM head trained to predict class
Chen [19]	0.694	Ensemble: 2x BERT + GPT-2 (PPL)
W. Huang [49]	0.683	Perplexity of GPT-2 trained on LLMs + SVM
Qin [103]	0.680*	Ensemble: BERTs + R-Drop
Binoculars 2	*0.671*	*Baseline Binoculars (Mistral-7B) [42]*
DetectLLM 1	*0.654*	*Baseline DetectLLM LRR (Mistral-7B) [120]*
Petropoulos [98]	0.641	RoBERTa embeddings + Bi-LSTM
Fast-DetectGPT 1	*0.638*	*Baseline Fast-DetectGPT (Mistral-7B) [10]*
Wu [136]	0.608	BERT embeddings + extra Transformer block
Text Length	*0.604*	*Baseline Text length*
Z. Lin [70]	0.565	T5 with LM head trained to predict class
Zhu [148]	0.555	Finetuned DeBERTa
PPMd CBC	*0.544*	*Baseline PPMd Compression-based Cosine [41, 114]*
Sun [122]	0.531	BERT embeddings + CNN
DetectLLM 2	*0.512*	*Baseline DetectLLM NPR (Mistral-7B) [120]*
Lei [65]	0.504	LoRA-trained ChatGLM
Fast-DetectGPT 2	*0.500*	*Baseline Fast-DetectGPT (Falcon-7B) [10]*
Liu [74]	0.497	Preplexity of pre-trained GPT-2
DetectGPT 1	*0.488*	*Baseline DetectGPT (Mistral-7B) [85]*
K. Huang [48]	0.480	Siamese DeBERTa
DetectLLM 3	*0.468*	*Baseline DetectLLM NPR (Falcon-7B) [120]*
Unmasking	*0.467*	*Baseline Authorship Unmasking [14, 58]*
Sheykhlan [116]	0.460	Ensemble: BERT, RoBERTa, and Electra
DetectLLM 4	*0.460*	*Baseline DetectLLM LRR (Falcon-7B) [120]*
DetectGPT 2	*0.439*	*Baseline DetectGPT (Falcon-7B) [85]*
Ostrower [94]		[No software submitted]

* Scores estimated due to run failures on some dataset variants.

LLM detection datasets. Some systems use engineered features like perplexity, properties of token distributions, or stylometrics (exclusively or in addition to BERT-embeddings) as classifier (Linear, XGBoost, LSTM) inputs. Most of these classification methods apply a posterior comparison of scores similar to how we use Binoculars, although some participants also train models to directly discriminate between the pairings. In some cases, participants also developed zero-shot methods and adapted LLMs directly for the detection task, often using LoRa.

Results

Table 8 shows the ranking scores of the best system submitted by each participating team and the baselines. In total, 10 teams surpassed all baselines. The overall best submission (by Tavan and Najafi; mean score of 0.924) finetunes Mistral and Llama2 models, combining them into an ensemble with the Binoculars baseline [42]. This approach beats the original baseline by 0.183 points, though there appears to be no general best strategy for AI detection. The top 5 systems are a mixture of zero-shot perplexity estimators and supervised blackbox classifiers based on BERT or even linear classifiers.

On the individual datasets, we see that almost all submissions perform quite well on non-obfuscated text (Roc-Auc > 0.9). We must therefore conclude that even the most advanced LLMs still exhibit obvious stylistic idiosyncrasies which make their texts easy to distinguish from human ones. However, none of the systems is entirely robust against (unexpected) obfuscations and particularly short text samples are a big challenge for all systems. Some systems did not produce any output on the short texts due to a programming problem. For the final evaluation, the missing values were filled with the corresponding mean values from all other systems. Affected systems are marked with * in Table 8.

A more detailed description and analysis of the submissions and the results can be found in the joint PAN and ELOQUENT task overview paper [15].

Acknowledgments. The work of Paolo Rosso, Damir Korenčić, and Berta Chulvi was in the framework of XAI-DisInfodemics: eXplainable AI for disinformation and conspiracy detection during infodemics (MICIN PLEC2021-007681), funded by MCIN/AEI / 10.13039/501100011033 and by the European Union NextGenerationEU / PRTR.

The work from Symanto has been partially funded by XAI-DisInfodemics: eXplainable AI for disinformation and conspiracy detection during infodemics (MICIN PLEC2021-007681), Pro^{2}Haters – Proactive Profiling of Hate Speech Spreaders (CDTi IDI-20210776), OBULEX - *OBservatorio del Uso de Lenguage sEXista en la red* (IVACE IMINOD/2022/106), and the ANDHI – ANomalous Diffusion of Harmful Information (CPP2021-008994) R&D grants.

The work of Janek Bevendorff, Matti Wiegmann, Maik Fröbe, Martin Potthast, and Benno Stein has been funded as part of the OpenWebSearch project by the European Commission (OpenWebSearch.eu, GA 101070014).

References

1. Abburi, H., Pudota, N., Veeramani, B., Bowen, E., Bhattacharya, S.: Team Deloitte at PAN: generative AI text detection. Working Notes of CLEF 2024, CEUR-WS.org (2024)
2. AI, M.: Kimi chatbot (2024). https://kimi.moonshot.cn. Accessed 31 May 2024
3. AI@Meta: Llama 3 model card (2024). https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md
4. Albladi, A., Seals, C.: Detection of Conspiracy vs. Critical Narratives and Their Elements using NLP. Working Notes of CLEF 2024, CEUR-WS.org (2024)
5. Almazrouei, E., et al.: The falcon series of open language models. arXiv [cs.CL] (2023)
6. Ansari, T., Ghazi, T., Alvi, F., Samad, A.: Decoding COVID-19 Narratives: Conspiracy or Critique? Working Notes of CLEF 2024, CEUR-WS.org (2024)
7. Ayele, A.A., Dinter, S., Belay, T.D., Asfaw, T.T., Yimam, S.M., Biemann, C.: The 5Js in Ethiopia: amharic hate speech data annotation using Toloka Crowdsourcing Platform. In: Proceedings of the 4th International Conference on Information and Communication Technology for Development for Africa (ICT4DA), pp. 114–120, Bahir Dar, Ethiopia (2022). https://ieeexplore.ieee.org/document/9971189
8. Ayele, A.A., Yimam, S.M., Belay, T.D., Asfaw, T., Biemann, C.: Exploring Amharic hate speech data collection and classification approaches. In: Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing (2023). https://aclanthology.org/2023.ranlp-1.6
9. Balasundaram, P., Swaminathan, K., Sampath, O., Km, P.: Oppositional Thinking Analysis: Conspiracy Theories vs Critical Thinking Narratives. Working Notes of CLEF 2024, CEUR-WS.org (2024)
10. Bao, G., Zhao, Y., Teng, Z., Yang, L., Zhang, Y.: Fast-DetectGPT: efficient zero-shot detection of machine-generated text via conditional probability curvature. arXiv [cs.CL] (2023)
11. Belchikov, A.: Russian language toxic comments (2019). https://www.kaggle.com/blackmoon/russian-language-toxic-comments. Accessed 14 Dec 2023
12. Bevendorff, J., et al.: Overview of PAN 2021: authorship verification, profiling hate speech spreaders on twitter, and style change detection. In: Candan, K.S., et al. (eds.) CLEF 2021. LNCS, vol. 12880, pp. 419–431. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-85251-1_26
13. Bevendorff, J., et al.: Overview of PAN 2020: authorship verification, celebrity profiling, profiling fake news spreaders on twitter, and style change detection. In: Arampatzis, A., et al. (eds.) CLEF 2020. LNCS, vol. 12260, pp. 372–383. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58219-7_25
14. Bevendorff, J., Stein, B., Hagen, M., Potthast, M.: Generalizing unmasking for short texts. In: Proceedings of the 2019 Conference of the North, pp. 654–659, Association for Computational Linguistics, Stroudsburg, PA, USA (2019). https://doi.org/10.18653/v1/n19-1068
15. Bevendorff, J., et al.: Overview of the "Voight-Kampff" generative AI authorship verification task at PAN and ELOQUENT 2024. Working Notes of CLEF 2024, CEUR Workshop Proceedings (2024)
16. Bobrovnyk, K.: Automated building and analysis of Ukrainian twitter corpus for toxic text detection. In: COLINS 2019. Volume II: Workshop (2019). https://ena.lpnu.ua:8443/server/api/core/bitstreams/c4c645c1-f465-4895-98dd-765f862cf186/content

17. Brundage, M., et al.: The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. CoRR **abs/1802.07228** (2018)
18. Cao, H., Han, Z., Ye, J., Liu, B., Han, Y.: Enhancing human-machine authorship discrimination in generative AI verification task with BERT and augmented data. Working Notes of CLEF 2024, CEUR-WS.org (2024)
19. Chen, J., Kong, L.: Integrating dual BERT models and causal language models for enhanced detection of machine-generated texts. Working Notes of CLEF 2024, CEUR-WS.org (2024)
20. Chen, Z., Han, Y., Yi, Y.: Team chen at PAN: integrating r-drop and pre-trained language model for multi-author writing style analysis. Working Notes of CLEF 2024, CEUR-WS.org (2024)
21. Chicco, D., Tötsch, N., Jurman, G.: The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Mining **14**(1), 13 (2021). https://doi.org/10.1186/s13040-021-00244-z. ISSN 1756-0381
22. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th ACL, ACL (2020). https://doi.org/10.18653/V1/2020.ACL-MAIN.747
23. Da San Martino, G., Barrón-Cedeño, A., Wachsmuth, H., Petrov, R., Nakov, P.: SemEval-2020 Task 11: detection of propaganda techniques in news articles. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 1377–1414, International Committee for Computational Linguistics, Barcelona (online) (2020). https://doi.org/10.18653/v1/2020.semeval-1.186. https://aclanthology.org/2020.semeval-1.186
24. Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., Nakov, P.: Fine-grained analysis of propaganda in news articles. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 5636–5646, Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1565. https://aclanthology.org/D19-1565
25. Damian, S., Herrera-Gonzalez, B., Vazquez-Santana, D., Calvo, H., Felipe-Riverón, E., Yáñez-Márquez, C.: DSVS at PAN 2024: ensemble approach of large language models for analyzing conspiracy theories against critical thinking narratives. Working Notes of CLEF 2024, CEUR-WS.org (2024)
26. Dementieva, D., Babakov, N., Panchenko, A.: Multiparadetox: extending text detoxification with parallel data to new languages. arXiv preprint arXiv:2404.02037 (2024)
27. Dementieva, D., et al.: RUSSE-2022: findings of the first Russian detoxification shared task based on parallel corpora. COMPUTATIONAL LINGUISTICS AND INTELLECTUAL TECHNOLOGIES (2022). https://api.semanticscholar.org/CorpusID:253169495
28. Dementieva, D., et al.: Crowdsourcing of parallel corpora: the case of style transfer for detoxification. In: Proceedings of the 2nd Crowd Science Workshop: Trust, Ethics, and Excellence in Crowdsourced Data Management at Scale co-located with 47th International Conference on Very Large Data Bases (VLDB 2021), CEUR Workshop Proceedings (2021). https://ceur-ws.org/Vol-2932/paper2.pdf
29. Douglas, K.M., Sutton, R.M.: What are conspiracy theories? A definitional approach to their correlates, consequences, and communication. Annu. Rev. Psychol. **74**(1), 271–298 (2023). https://doi.org/10.1146/annurev-psych-032420-031329

30. Espinosa, D., Sidorov, G., Ricárdez-Vázquez, E.: Using BERT to identify conspiracy theories. Working Notes of CLEF 2024, CEUR-WS.org (2024)
31. Feng, F., Yang, Y., Cer, D., Arivazhagan, N., Wang, W.: Language-agnostic BERT sentence embedding. In: Proceedings of the 60th ACL, ACL (2022). https://doi.org/10.18653/V1/2022.ACL-LONG.62
32. Fröbe, M., et al.: Continuous integration for reproducible shared tasks with TIRA.io. In: Kamps, J. (ed.) ECIR 2023. LNCS, vol. 13982, pp. 236–241. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-28241-6_20
33. Gambini, M., Tardelli, S., Tesconi, M.: The anatomy of conspiracy theorists: unveiling traits using a comprehensive twitter dataset. Comput. Commun. **217**, 25–40 (2024). https://doi.org/10.1016/j.comcom.2024.01.027
34. Gangopadhyay, S., Khan, M., Jabeen, H.: HybridDetox: Combining supervised and unsupervised methods for effective multilingual text detoxification. Working Notes of CLEF 2024, CEUR-WS.org (2024)
35. Giachanou, A., Ghanem, B., Rosso, P.: Detection of conspiracy propagators using psycho-linguistic characteristics. J. Inf. Sci. **49**(1), 3–17 (2023). https://doi.org/10.1177/0165551520985486
36. Gómez-Romero, J., González-Silot, S., Montoro-Montarroso, A., Molina-Solana, M., Martínez Cámara, E.: Detection of conspiracy-related messages in Telegram with anonymized named entities. Working Notes of CLEF 2024, CEUR-WS.org (2024)
37. Gritsai, G., Boyeva, G., Grabovoy, A.: Team ap-team at PAN: LLM Adapters for Various Datasets. Working Notes of CLEF 2024, CEUR-WS.org (2024)
38. Guo, L., Yang, W., Ma, L., Ruan, J.: BLGAV: generative AI author verification model based on BERT and BiLSTM. Working Notes of CLEF 2024, CEUR-WS.org (2024)
39. Guo, M., Han, Z., Chen, H., Peng, J.: A machine-generated text detection model based on text multi-feature fusion. Working Notes of CLEF 2024, CEUR-WS.org (2024)
40. Haddad, H., Mulki, H., Oueslati, A.: T-HSAB: a Tunisian hate speech and abusive dataset. In: Smaïli, K. (ed.) ICALP 2019. CCIS, vol. 1108, pp. 251–263. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32959-4_18
41. Halvani, O., Winter, C., Graner, L.: On the usefulness of compression models for authorship verification. In: Proceedings of the 12th International Conference on Availability, Reliability and Security, vol. Part F1305, ACM, New York, NY, USA (2017). https://doi.org/10.1145/3098954.3104050. ISBN 9781450352574
42. Hans, A., et al.: Spotting LLMs with binoculars: zero-shot detection of machine-generated text. arXiv [cs.CL] (2024)
43. Hong, J., Lee, N., Thorne, J.: ORPO: monolithic preference optimization without reference model. CoRR **abs/2403.07691** (2024). https://doi.org/10.48550/ARXIV.2403.07691
44. Hu, Q., Han, Z., Peng, J., Guo, M., Liu, C.: An oppositional thinking analysis method using BERT-based model with BiGRU. Working Notes of CLEF 2024, CEUR-WS.org (2024)
45. Huang, B., Zhong, C., Yan, K., Han, Y.: Author authentication of generative AI based on BERT by regularization method. Working Notes of CLEF 2024, CEUR-WS.org (2024)
46. Huang, J., Chen, Y., Luo, M., Li, Y.: Generative AI authorship verification of tri-sentence analysis base on the bert model. Working Notes of CLEF 2024, CEUR-WS.org (2024)

47. Huang, J., Han, Z., Zhu, R., Guo, M., Sun, K.: Conspiracy theory text classification based on CT-BERT and BETO models. Working Notes of CLEF 2024, CEUR-WS.org (2024)
48. Huang, K., Qi, H., Yan, K.: Voight-Kampff generative AI authorship verification based on contrastive learning and domain adaptation. Working Notes of CLEF 2024, CEUR-WS.org (2024)
49. Huang, W., Grieve, J.: Authorial language models for AI authorship verification. Working Notes of CLEF 2024, CEUR-WS.org (2024)
50. Huang, Z., Kong, L.: Team huangzhijian at PAN: DeBERTa-v3 with R-drop regularization for multi-author writing style analysis. Working Notes of CLEF 2024, CEUR-WS.org (2024)
51. Huertas-García, Á., Martí-González, C., Muñoz, J., Ambite, E.: Small language models and large language models in oppositional thinking analysis: capabilities and biases and challenges. Working Notes of CLEF 2024, CEUR-WS.org (2024)
52. Jiang, A.Q., et al.: Mistral 7B. arXiv [cs.CL] (2023)
53. Jigsaw: Toxic comment classification challenge (2017). https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge. Accessed 18 Mar 2024
54. Karlgren, J., Dürlich, L., Gogoulou, E., Guillou, L., Nivre, J., Sahlgren, M., Talman, A.: ELOQUENT CLEF shared tasks for evaluation of generative language model quality. In: Goharian, N., et al. (eds.) ECIR 2024. LNCS, vol. 14612, pp. 459–465. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-56069-9_63
55. Kestemont, M., et al.: Overview of the author identification task at PAN 2018: cross-domain authorship attribution and style change detection. In: Working Notes of CLEF 2018, CEUR-WS.org (2018)
56. Khan, A., Rai, M., Khan, K., Shah, S., Alvi, F., Samad, A.: Team gladiators at PAN: improving author identification: a comparative analysis of pre-trained transformers for multi-author classification. Working Notes of CLEF 2024, CEUR-WS.org (2024)
57. King, J., Baffour, P., Crossley, S., Holbrook, R., Demkin, M.: Llm – detect ai generated text (2023). https://kaggle.com/competitions/llm-detect-ai-generated-text
58. Koppel, M., Schler, J.: Authorship verification as a one-class classification problem. In: Twenty-first International Conference on Machine Learning - ICML 2004, pp. 489–495, ACM Press, New York (2004). https://doi.org/10.1145/1015330.1015448. ISBN 9781581138283
59. Koppel, M., Winter, Y.: Determining if two documents are written by the same author. J. Am. Soc. Inf. Sci. **65**(1), 178–187 (2014)
60. Korenčić, D., Chulvi, B., Bonet, X., Mariona, T., Toselli, A., Rosso, P.: What distinguishes conspiracy from critical narratives? A computational analysis of oppositional discourse. expert systems. Expert Syst. (2024)
61. Korenčić, D., Chulvi, B., Bonet Casals, X., Taulé, M., Rosso, P., Rangel, F.: Overview of the oppositional thinking analysis pan task at clef 2024. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
62. Korenčić, D., Grubišić, I., Toselli, A.H., Chulvi, B., Rosso, P.: Tackling Covid-19 conspiracies on twitter using BERT ensembles, GPT-3 augmentation, and graph NNs. In: Working Notes Proceedings of the MediaEval 2022 Workshop Bergen, Norway and Online (2023). https://2022.multimediaeval.com/paper8969.pdf
63. Księżniak, E., Węcel, K., Sawiński, M.: Team OpenFact at PAN 2024: Fine-tuning BERT models with stylometric enhancements. Working Notes of CLEF 2024, CEUR-WS.org (2024)

64. Kumar, S., Balachandran, V., Njoo, L., Anastasopoulos, A., Tsvetkov, Y.: Language generation models can cause harm: So what can we do about it? an actionable survey. CoRR **abs/2210.07700** (2022)
65. Lei, H., Liu, X., Niu, G., Zhou, Y., Zhou, Y.: Generative AI authorship verification based on ChatGLM. Working Notes of CLEF 2024, CEUR-WS.org (2024)
66. Liang, X., Lei, H.: Team lxflcl66666 at PAN: fine-tuned reasoning for writing style analysis. Working Notes of CLEF 2024, CEUR-WS.org (2024)
67. Liang, X., et al.: R-drop: regularized dropout for neural networks. In: 34th Annual Conference on Neural Information Processing Systems 2021, NeurIPS (2021)
68. Lin, T., Wu, Y., Lee, L.: Team NYCU-NLP at PAN 2024: integrating transformers with similarity adjustments for multi-author writing style analysis. Working Notes of CLEF 2024, CEUR-WS.org (2024)
69. Lin, Z., et al.: A verifying generative text authorship model with regularized dropout. Working Notes of CLEF 2024, CEUR-WS.org (2024)
70. Lin, Z., Li, Y., Huang, J.: Voight-Kampff generative AI authorship verification based on T5. Working Notes of CLEF 2024, CEUR-WS.org (2024)
71. Liu, B., Han, Z., Cao, H.: An approach to classifying conspiratorial and critical public health narratives. Working Notes of CLEF 2024, CEUR-WS.org (2024)
72. Liu, C., Han, Z., Chen, H., Hu, Q.: Team liuc0757 at PAN: a writing style embedding method based on contrastive learning for multi-author writing style analysis. Working Notes of CLEF 2024, CEUR-WS.org (2024)
73. Liu, X., Chen, H., Lv, J.: Team foshan-university-of-guangdong at PAN: adaptive entropy-based stability-plasticity for multi-author writing style analysis. Working Notes of CLEF 2024, CEUR-WS.org (2024)
74. Liu, X., Kong, L.: AI text detection method based on perplexity features with strided sliding window. Working Notes of CLEF 2024, CEUR-WS.org (2024)
75. Logacheva, V., et al.: ParaDetox: detoxification with parallel data. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 6804–6818, Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.acl-long.469. https://aclanthology.org/2022.acl-long.469
76. Lorenz, L., Aygüler, F.Z., Schlatt, F., Mirzakhmedova, N.: BaselineAvengers at PAN 2024: often-forgotten baselines for LLM-generated text detection. Working Notes of CLEF 2024, CEUR-WS.org (2024)
77. Lu, J., Xu, B., Zhang, X., Min, C., Yang, L., Lin, H.: Facilitating fine-grained detection of Chinese toxic language: hierarchical taxonomy, resources, and benchmarks. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, pp. 16235–16250 (2023). https://aclanthology.org/2023.acl-long.898
78. Luo, Z., Luo, M., Wang, A.: Multilingual text detoxification using google cloud translation and post-processing. Working Notes of CLEF 2024, CEUR-WS.org (2024)
79. Lv, J., Han, Y., Kong, L.: Meta-contrastive learning for generative AI authorship verification. Working Notes of CLEF 2024, CEUR-WS.org (2024)
80. Lv, J., Yi, Y., Qi, H.: Team Fosu-stu at PAN: supervised fine-tuning of large language models for multi author writing style analysis. Working Notes of CLEF 2024, CEUR-WS.org (2024)
81. Mahesh, S., Divakaran, S., Girish, K., Lakshmaiah, S.: Binary battle: leveraging ML and TL models to distinguish between conspiracy theories and critical thinking. Working Notes of CLEF 2024, CEUR-WS.org (2024)

82. Mandl, T., et al.: Overview of the hasoc track at fire 2019: hate speech and offensive content identification in Indo-European languages. In: Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation, pp. 14–17, FIRE 2019, ACM (2019). https://doi.org/10.1145/3368567.3368584. ISBN 9781450377508
83. Mathet, Y., Widlöcher, A., Métivier, J.P.: The unified and holistic method gamma for inter-annotator agreement measure and alignment. Comput. Linguist. **41**(3), 437–479 (2015). https://doi.org/10.1162/COLI_a_00227. ISSN 0891-2017
84. Miralles, P., Martín, A., Camacho, D.: Ensembling normalized log probabilities. Working Notes of CLEF 2024, CEUR-WS.org (2024)
85. Mitchell, E., Lee, Y., Khazatsky, A., Manning, C.D., Finn, C.: DetectGPT: zero-shot machine-generated text detection using probability curvature. In: International Conference on Machine Learning, vol. 202, pp. 24950–24962 (2023). https://doi.org/10.48550/arXiv.2301.11305
86. MTS.AI: Cotype: Generative AI solutions (2022). https://mts.ai. Accessed 31 May 2024
87. Mubarak, H., Darwish, K., Magdy, W., Elsayed, T., Al-Khalifa, H.: Overview of osact4 Arabic offensive language detection shared task. In: Proceedings of the 4th Workshop on Open-source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, pp. 48–52 (2020)
88. Muennighoff, N., et al.: Crosslingual generalization through multitask finetuning. In: Proceedings of the 61st ACL, ACL (2023). https://doi.org/10.18653/V1/2023.ACL-LONG.891
89. Mulki, H., Ghanem, B.: Let-mi: An Arabic levantine twitter dataset for misogynistic language. In: Habash, N., (eds.) et al Proceedings of the Sixth Arabic Natural Language Processing Workshop, pp. 154–163, Association for Computational Linguistics, Kyiv, Ukraine (Virtual) (2021). https://aclanthology.org/2021.wanlp-1.16
90. Mulki, H., Haddad, H., Ali, C.B., Alshabani, H.: L-hsab: a levantine twitter dataset for hate speech and abusive language. In: Proceedings of the Third Workshop on Abusive Language Online, pp. 111–118 (2019)
91. Najafi, M., Tavan, E., Colreavy, S.: Marsan at PAN 2024 TextDetox: ToxiCleanse RL and paving the way for toxicity-free online discourse. Working Notes of CLEF 2024, CEUR-WS.org (2024)
92. OpenAI: Chatgpt: Optimizing language models for dialogue (2022). https://openai.com/blog/chatgpt Accessed 31 May 2024
93. Osipenko, M., Korchagin, M., Toleugazinov, A., Egorov, S., Udobang, J.: Fancy transformers at PAN 2024 TextDetox: surpassing the baselines. Working Notes of CLEF 2024, CEUR-WS.org (2024)
94. Ostrower, B., Wessell, J., Bindal, A.: AI authorship verification: an ensembled approach. Working Notes of CLEF 2024, CEUR-WS.org (2024)
95. Peng, J., et al.: A multilingual text detoxification method based on few-shot learning and CO-STAR framework. Working Notes of CLEF 2024, CEUR-WS.org (2024)
96. Pereira-Kohatsu, J.C., Sánchez, L.Q., Liberatore, F., Camacho-Collados, M.: Detecting and monitoring hate speech in twitter. Sensors **19**(21), 4654 (2019). https://doi.org/10.3390/S19214654
97. Pérez, J.M., Furman, D.A., Alonso Alemany, L., Luque, F.M.: RoBERTuito: a pre-trained language model for social media text in Spanish. In: Proceedings of the 13th LREC, ELRA (2022). https://aclanthology.org/2022.lrec-1.785

98. Petropoulos, P., Petropoulos, V.: RoBERTa and Bi-LSTM for human vs AI generated text detection. Working Notes of CLEF 2024, CEUR-WS.org (2024)
99. Pletenev, S.: Memu_pro_kotow at PAN 2024 TextDetox: uncensored Llama3 helps to censor better. Working Notes of CLEF 2024, CEUR-WS.org (2024)
100. Pogorelov, K., Schroeder, D.T., Brenner, S., Langguth, J.: FakeNews: corona virus and conspiracies multimedia analysis task at MediaEval 2021. In: Working Notes Proceedings of the MediaEval 2021 Workshop Bergen, Norway and Online (2021)
101. Pogorelov, K., Schroeder, D.T., Brenner, S., Maulana, A., Langguth, J.: Combining tweets and connections graph for fakenews detection at mediaeval 2022. In: Proceedings of the MediaEval 2022 Workshop, Bergen, Norway and Online, 12-13 January 2023. (2023)
102. Protasov, V.: PAN 2024 Multilingual TextDetox: exploring cross-lingual transfer in case of large language models. Working Notes of CLEF 2024, CEUR-WS.org (2024)
103. Qin, R., Qi, H., Yi, Y.: A model fusion approach for generative AI authorship verification. Working Notes of CLEF 2024, CEUR-WS.org (2024)
104. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv [cs.LG] (2019)
105. Řehulka, E., Šuppa, M.: RAG meets detox: enhancing text detoxification using open-source large language models with retrieval augmented generation. Working Notes of CLEF 2024, CEUR-WS.org (2024)
106. Risch, J., Stoll, A., Wilms, L., Wiegand, M.: Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In: Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments, pp. 1–12, Duesseldorf, Germany (2021)
107. Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., Wojatzki, M.: Measuring the reliability of hate speech annotations: the case of the european refugee crisis. In: Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication, Bochumer Linguistische Arbeitsberichte, vol. 17, pp. 6–.9, Bochum, Germany (2016)
108. Rosso, P., Rangel, F., Potthast, M., Stamatatos, E., Tschuggnall, M., Stein, B.: Overview of PAN'16. In: Fuhr, N., et al. (eds.) CLEF 2016. LNCS, vol. 9822, pp. 332–350. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44564-9_28
109. Ruffo, G., Semeraro, A., Giachanou, A., Rosso, P.: Studying fake news spreading, polarisation dynamics, and manipulation by bots: a tale of networks and language. Comput. Sci. Rev. **47**, 100531 (2023). https://doi.org/10.1016/j.cosrev.2022.100531. https://www.sciencedirect.com/science/article/pii/S157401372200065X. ISSN 1574-0137
110. Rykov, E., Zaytsev, K., Anisimov, I., Voronin, A.: SmurfCat at PAN TexDetox 2024: alignment of multilingual transformers for text detoxification. Working Notes of CLEF 2024, CEUR-WS.org (2024)
111. Sahitaj, A., Sahitaj, P., Mohtaj, S., Möller, S., Schmitt, V.: Towards a computational framework for distinguishing critical and conspiratorial texts by elaborating on the context and argumentation with LLMs. Working Notes of CLEF 2024, CEUR-WS.org (2024)
112. Sánchez-Hermosilla, I., Panizo Lledot, A., Camacho, D.: A study on NLP model ensembles and data augmentation techniques for separating critical thinking from conspiracy theories in English texts. Working Notes of CLEF 2024, CEUR-WS.org (2024)

113. Sanjesh, R., Mangai, A.: Team riyasanjesh at PAN: multi-feature with CNN and Bi-LSTM neural network approach to style change detection. Working Notes of CLEF 2024, CEUR-WS.org (2024)
114. Sculley, D., Brodley, C.E.: Compression and machine learning: A new perspective on feature space vectors. In: Data Compression Conference (DCC 2006), pp. 332–341, IEEE (2006). https://doi.org/10.1109/dcc.2006.13. ISBN 9780769525457, ISSN 1068-0314,2375-0359
115. Semiletov, A.: Toxic Russian comments: labelled comments from the popular Russian social network (2020). https://www.kaggle.com/alexandersemiletov/toxic-russian-comments. Accessed 14 Dec 2023
116. Sheykhlan, M., Abdoljabbar, S., Mahmoudabad, M.: Team karami-kheiri at PAN: enhancing machine-generated text detection with ensemble learning based on transformer models. Working Notes of CLEF 2024, CEUR-WS.org (2024)
117. Sheykhlan, M., Abdoljabbar, S., Mahmoudabad, M.: Team karami-sh at PAN: transformer-based ensemble learning for multi-author writing style analysis. Working Notes of CLEF 2024, CEUR-WS.org (2024)
118. Stamatatos, E., et al.: Overview of the authorship verification task at PAN 2022. In: CLEF 2022 Labs and Workshops, CEUR-WS.org (2022)
119. Stamatatos, E., Potthast, M., Rangel, F., Rosso, P., Stein, B.: Overview of the PAN/CLEF 2015 evaluation lab. In: Mothe, J., et al. (eds.) CLEF 2015. LNCS, vol. 9283, pp. 518–538. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24027-5_49
120. Su, J., Zhuo, T.Y., Wang, D., Nakov, P.: DetectLLM: leveraging log rank information for zero-shot detection of machine-generated text. arXiv [cs.CL] (2023)
121. Su, Y., Lan, T., Wang, Y., Yogatama, D., Kong, L., Collier, N.: A contrastive framework for neural text generation. arXiv [cs.CL] (2022)
122. Sun, G., Yang, W., Ma, L.: BCAV: a generative ai author verification model based on the integration of Bert and CNN. Working Notes of CLEF 2024, CEUR-WS.org (2024)
123. Sushko, N.: PAN 2024 multilingual TextDetox: exploring different regimes for synthetic data training for multilingual text detoxification. Working Notes of CLEF 2024, CEUR-WS.org (2024)
124. Taulé, M., Nofre, M., Bargiela, V., Bonet, X.: Newscom-tox: a corpus of comments on news articles annotated for toxicity in spanish. LREC (2024)
125. Tavan, E., Najafi, M.: Marsan at PAN: BinocularLLM and fusing binoculars' insight with the proficiency of large language models for cutting-edge machine-generated text detection. Working Notes of CLEF 2024, CEUR-WS.org (2024)
126. Tian, Y., et al.: Multiscale positive-unlabeled detection of AI-generated texts. CoRR **abs/2305.18149** (2023). https://doi.org/10.48550/ARXIV.2305.18149
127. Tschuggnall, M., et al.: Overview of the author identification task at PAN 2017: style breach detection and author clustering. In: CLEF 2017 Labs and Workshops (2017)
128. Tulbure, A., Coll Ardanuy, M.: Conspiracy vs critical thinking using an ensemble of transformers with data augmentation techniques. Working Notes of CLEF 2024, CEUR-WS.org (2024)
129. Valdez-Valenzuela, A., Gómez-Adorno, H.: Team iimasnlp at PAN: leveraging graph neural networks and large language models for generative AI authorship verification. Working Notes of CLEF 2024, CEUR-WS.org (2024)
130. Vallecillo-Rodríguez, M., Martín-Valdivia, A.M.: SINAI at PAN 2024 TextDetox: application of tree of thought strategy in large language models for multilingual text detoxification. Working Notes of CLEF 2024, CEUR-WS.org (2024)

131. Vallecillo-Rodríguez, M., Martín-Valdivia, M., Montejo-Ráez, A.: SINAI at PAN 2024 oppositional thinking analysis: exploring the fine-tuning performance of LLMs. Working Notes of CLEF 2024, CEUR-WS.org (2024)
132. Weimer, A.M., et al.: The (in-)consistency of literary concepts. operationalising, annotating and detecting literary comment. J. Comput. Literary Stud. **1**(1) (2022). https://doi.org/10.48694/jcls.90, https://jcls.io/article/id/90/. ISSN 2940-1348
133. Wiegand, M., Siegel, M., Ruppenhofer, J.: Overview of the GermEval 2018 shared task on the identification of offensive language (2018)
134. Wu, B., Han, Y., Yan, K., Qi, H.: Team baker at PAN: enhancing writing style change detection with virtual softmax. Working Notes of CLEF 2024, CEUR-WS.org (2024)
135. Wu, Q., Kong, L., Ye, Z.: Team bingezzzleep at PAN: a writing style change analysis model based on RoBERTa encoding and contrastive learning for multi-author writing style analysis. Working Notes of CLEF 2024, CEUR-WS.org (2024)
136. Wu, Z., Yang, W., Ma, L., Zhao, Z.: BertT: a hybrid neural network model for generative AI authorship verification. Working Notes of CLEF 2024, CEUR-WS.org (2024)
137. Xue, L., et al.: mt5: a massively multilingual pre-trained text-to-text transformer. In: Proceedings of the NAACL-HLT 2021, ACL. https://doi.org/10.18653/V1/2021.NAACL-MAIN.41
138. Yadagiri, A., Kalita, D., Ranjan, A., Bostan, A., Toppo, P., Pakray, P.: Team cnlp-nits-pp at PAN: leveraging BERT for accurate authorship verification: a novel approach to textual attribution. Working Notes of CLEF 2024, CEUR-WS.org (2024)
139. Ye, Z., Zhong, Y., Huang, C., Kong, L.: Team no-999 at PAN: continual transfer learning with progress prompt for multi-author writing style analysis. Working Notes of CLEF 2024, CEUR-WS.org (2024)
140. Ye, Z., Zhong, Y., Huang, Z., Kong, L.: Token prediction as implicit classification for generative AI authorship verification. Working Notes of CLEF 2024, CEUR-WS.org (2024)
141. Zangerle, E., Mayerl, M., , Potthast, M., Stein, B.: Overview of the style change detection task at PAN 2021. In: Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.) CLEF 2021 Labs and Workshops, CEUR-WS.org (2021)
142. Zangerle, E., Mayerl, M., , Potthast, M., Stein, B.: Overview of the style change detection task at PAN 2022. In: CLEF 2022 Labs and Workshops, CEUR-WS.org (2022)
143. Zangerle, E., Mayerl, M., , Potthast, M., Stein, B.: Overview of the style change detection task at PAN 2023. In: CLEF 2023 Labs and Workshops, CEUR-WS.org (2023)
144. Zangerle, E., Mayerl, M., Potthast, M., Stein, B.: Overview of the multi-author writing style analysis task at PAN 2024. Working Notes of CLEF 2024, CEUR-WS.org (2024)
145. Zangerle, E., Mayerl, M., Specht, G., Potthast, M., Stein, B.: Overview of the style change detection task at PAN 2020. In: CLEF 2020 Labs and Workshops (2020)
146. Zangerle, E., Tschuggnall, M., Specht, G., Stein, B., Potthast, M.: Overview of the style change detection task at PAN 2019. In: CLEF 2019 Labs and Workshops (2019)

147. Zeng, Z., et al.: A conspiracy theory text detection method based on RoBERTa and XLM-RoBERTa models. In: Working Notes of CLEF 2024, CEUR-WS.org (2024)
148. Zhu, Y., Kong, L.: AI Authorship verification based on deberta model. Working Notes of CLEF 2024, CEUR-WS.org (2024)
149. Zinkovich, V., Karpukhin, S., Kurdiukov, N., Tikhomirov, P.: nlp_enjoyers at multilingual textual detoxification (CLEF-2024. Working Notes of CLEF 2024, CEUR-WS.org (2024)
150. Zrnić, L.: Conspiracy theory detection using transformers with multi-task and multilingual approaches. Working Notes of CLEF 2024, CEUR-WS.org (2024)

Overview of QuantumCLEF 2024: The Quantum Computing Challenge for Information Retrieval and Recommender Systems at CLEF

Andrea Pasin[1(✉)], Maurizio Ferrari Dacrema[2], Paolo Cremonesi[2], and Nicola Ferro[1]

[1] University of Padua, Padua, Italy
andrea.pasin.1@phd.unipd.it, nicola.ferro@unipd.it
[2] Politecnico di Milano, Milan, Italy
{maurizio.ferrari,paolo.cremonesi}@polimi.it

Abstract. *Quantum Computing (QC)* is an innovative research field that has gathered the interest of many researchers in the last few years. In fact, it is believed that QC could potentially revolutionize the way we solve very complex problems by dramatically decreasing the time required to solve them. Even though QC is still in its early stages of development, it is already possible to tackle some problems by means of quantum computers and to start catching a glimpse of its potential. Therefore, the aim of the QuantumCLEF lab is to raise awareness about QC and to develop and evaluate new QC algorithms to solve challenges that can be encountered when implementing *Information Retrieval (IR)* and *Recommender Systems (RS)* systems. Furthermore, this lab represents a good opportunity to engage with QC technologies which are typically not easily accessible.

In this work, we present an overview of the first edition of QuantumCLEF, a lab that focuses on the application of *Quantum Annealing (QA)*, a specific QC paradigm, to solve two tasks: Feature Selection for IR and RS systems, and Clustering for IR systems. There have been a total of 26 teams who registered for this lab and eventually 7 teams managed to successfully submit their runs following the lab guidelines. Due to the novelty of the topics, participants have been provided with many examples and comprehensive materials that allowed them to understand how QA works and how to program quantum annealers.

1 Introduction

Information Retrieval (IR) and *Recommender Systems (RS)* systems have been studied and improved for several years. Nowadays, these systems need to face very complex challenges such as applying computationally expensive methods to huge amounts of data that are constantly being produced.

To solve this issue, researchers are now investigating *Quantum Computing (QC)*, an emerging computing paradigm that has the potential to revolutionize

L. Goeuriot et al. (Eds.): CLEF 2024, LNCS 14959, pp. 260–282, 2024.
https://doi.org/10.1007/978-3-031-71908-0_12

the way we currently solve problems. QC is not only about a new technology that can be used in place of traditional hardware, but it also represents a paradigm shift that allows to view and solve problems from a new perspective exploiting quantum physics principles. Thanks to principles such as superposition and entanglement, quantum computers can theoretically explore exponentially larger problem spaces with respect to traditional computers considering devices with the same number of quantum bits (qubits) and traditional bits respectively.

In recent years, quantum computers have started to become more robust, powerful, and accessible. This has allowed researchers and practitioners to start exploring the application of QC to practical problems. However, QC is still in its infancy and there are several limitations yet to overcome, most of which concerning the hardware. In fact, qubits are very delicate and must be completely isolated from the environment since any interferences or noises (e.g., electromagnetic interferences, thermal fluctuations) could impact their state, thus breaking the computation. On the other hand, traditional systems have been developed for decades and they represent more robust alternatives.

In this exciting and innovative context, it is natural to wonder whether it is possible to apply QC to solve some of the complex tasks that are faced by IR and RS systems. For this reason, we decided to start a new CLEF lab called QuantumCLEF [20,21] which focuses on the study, development, and evaluation of QC algorithms for IR and RS. This lab has 4 main objectives:

- develop new QC algorithms for IR and RS and evaluate them, comparing the results (efficiency and effectiveness) with traditional approaches;
- gather all resources and data for future researchers to compare their results with the ones achieved during the lab;
- allow participants to learn more about QC through comprehensive materials and to use real quantum computers, which are still not easily accessible to the public;
- raise the awareness of the potential of QC and form a new research community around this new field.

In this paper, we present the overview of the first edition of QuantumCLEF held in 2024. This edition has focused on the usage of *Quantum Annealing (QA)*, a specific QC paradigm that can be used to tackle optimization problems. We have granted participants access to the state-of-the-art QA devices (quantum annealers) produced by D-Wave, one of the leading companies in this sector. The QA paradigm is easier to understand with respect to the Universal Gate-Based paradigm. Furthermore, D-Wave provides several tools and libraries to program quantum annealers without requiring a very deep knowledge of the quantum physics governing these devices.

This QuantumCLEF edition was composed of two main tasks:

- **Task 1**: Feature Selection for IR and RS;
- **Task 2**: Clustering for IR.

Participants were asked to develop their own algorithms to solve the tasks using both QA and *Simulated Annealing (SA)*, a well-known optimization approach similar to QA but without any quantum effects and therefore can be run on classical devices. Due to the novelty of the topics, comprehensive materials (i.e., videos, slides, and examples) were provided to the participants to lower the entry barrier and to allow them to understand how QA works and how to program quantum annealers. An ad-hoc infrastructure has been created to grant participants access to real quantum annealers while also easing the workflow and enhancing reproducibility. In total 26 teams participated in our tasks, 7 of which actively participated and submitted their runs. More specifically, 6 teams managed to successfully submit their runs for Task 1 while 1 team managed to submit for Task 2. The results show that approaches that use QA or *Hybrid (H)* methods are as effective as SA and traditional approaches while being generally more efficient.

The paper is organized as follows: Sect. 2 discusses related works; Sect. 3 presents the tasks of the QuantumCLEF 2024 lab while Sect. 4.1 introduces the lab's setup and the design and implementation of our ad-hoc infrastructure; Sect. 5 shows and discusses the results achieved by the participants; finally, Sect. 6 draws some conclusions and outlooks some future work.

2 Related Works

2.1 Background on Quantum and Simulated Annealing

We provide here a brief introduction to QA and to *Simulated Annealing (SA)*, a traditional optimization algorithm that does not take advantage of quantum technologies.

Quantum Annealing. QA is a QC paradigm that is based on special-purpose devices (quantum annealers) able to tackle optimization problems with a certain structure. The basic idea of a quantum annealer is to represent a problem as the energy of a physical system and then leverage quantum-mechanical phenomena, e.g., superposition and entanglement, to let the system find a state of minimal energy, which corresponds to the solution of the original problem.

To use quantum annealers, one needs to formulate the optimization problem as a minimization one using the *Quadratic Unconstrained Binary Optimization (QUBO)* formulation [14], a well-known optimization technique. QUBO is defined as:

$$\min \quad y = x^T Q x \tag{1}$$

where x is a vector of binary decision variables, and Q is a matrix of constant values representing the problem we wish to solve. Then, a further step called *minor embedding* is required to map the general mathematical formulation into the physical quantum annealer hardware, accounting for the limited number of qubits and the physical connections between them. Each quantum annealer or *Quantum Processing Unit (QPU)* has, in fact, its own architecture, which can be seen as a graph: each vertex represents a qubit, and each edge represents an interaction between two qubits. Therefore, minor embedding involves choosing which physical qubits represent the decision variables. If the QUBO problem does not fit directly in the QPU, for example because a decision variable is connected to more variables than the available physical connections between qubits, multiple connected qubits will be used to represent one decision variable and the connections to the other variables will be split between them. Due to this the number of qubits required to solve a problem on a quantum annealer may be much higher than the number of its decision variables. Minor embedding is a complex task in itself and a NP-hard problem, which can be solved relying on some heuristic methods [8]. If the problem does not fit on the QPU, D-Wave provides *Hybrid (H)* approaches that are able to automatically handle large problems using intelligent techniques to split them and solve them using both traditional methods and QA methods. By splitting problems into sub-problems it will be possible to make them fit inside the QPU of quantum annealers.

Occasionally, it might be necessary to add constraints to the problems. This can be done by means of penalties $\mathrm{P}(x)$ [28], which penalize solutions that do not meet the specified constraints. These penalties are then added to the original cost function y to achieve the final formulation as follows:

$$\min \quad \mathrm{C}(x) = y + \mathrm{P}(x)\ . \tag{2}$$

Penalties can be controlled through hyperparameters to manage their influence with respect to the given formulation.

To sum up, using a quantum annealer requires several stages [28]:

1. **Formulation**: find a way to express the desired algorithm as an optimization problem by leveraging the QUBO framework and compute the actual QUBO matrix Q;
2. **Embedding**: generate the minor embedding of the QUBO for the quantum annealer hardware;
3. **Data Transfer**: transfer the problem and the embedding on the global network to the data center that hosts the quantum annealer;
4. **Annealing**: run the quantum annealer itself. This phase is composed by several stages such as programming the QPU, sampling a solution, and then reading the solution. This is an inherently stochastic process. Therefore, it is usually run a large number of times (hundreds) in which several samples are returned, each one resembling a possible solution to the considered problem. The solutions must then be checked for their feasibility, and then the best one among them (i.e., the optimal one according to the objective function) is usually considered the final solution to the submitted problem.

Generally, once a QUBO problem has been embedded and sent to the quantum annealer, it can be solved in a few milliseconds.

Simulated Annealing. SA is a consolidated meta-heuristic that can be run on traditional hardware [6,26]. It is a probabilistic algorithm that can be used to find the global minimum of a given cost function, even in the presence of many local minima. It is based on an iterative process that starts from an initial solution and tries to improve it by randomly perturbing it. The cost function is represented by the QUBO problem formulation, similar to what would be used for QA. In SA, there is no minor embedding phase since the problem is directly solved on a traditional machine.

We underline that SA is an optimization algorithm different from QA, it is not a simulation of QA on a traditional machine, and, therefore these two algorithms are not equivalent. However, SA can be used for benchmarking purposes to show how well QA performs with respect to a traditional hardware counterpart.

The access to quantum annealers in QuantumCLEF is limited to ensure a fair distribution of resources. Therefore, SA can also be used to perform initial experiments to assess a QUBO formulation feasibility without affecting the available quota in the quantum environment.

2.2 Related Challenges

In the context of CLEF, there have not been other challenges involving the application and evaluation of QC. However, since QC technologies are starting to become more available and robust, it is necessary to raise awareness about their potential and to learn how these technologies can be used to possibly improve the current state-of-the-art IR and RS systems.

Outside CLEF, we are not aware of other challenges or shared tasks that have been done in the past involving the use of QC. There are some other challenges starting off this year offered by big-tech companies such as IBM[1] and Google[2]. These challenges involve the development of QC algorithms which will be executed on quantum computers to solve some practical real-world challenges. There has also been a Quantum Computing challenge in 2016 organized by Microsoft[3], which however used simulators for Language-Integrated Quantum Operations and not real quantum computers.

3 Tasks

QuantumCLEF 2024, which was initially presented in a paper at CLEF 2023 [20], addresses two different tasks involving computationally intensive problems that are closely related to the Information Access field: Feature Selection and Clustering. The main goals for each task are:

- finding one or more possible QUBO formulations of the problem;
- evaluating the QA approach compared to a corresponding traditional approach to assess both its efficiency and its effectiveness.

[1] https://challenges.quantum.ibm.com/2024.
[2] https://www.xprize.org/prizes/qc-apps.
[3] https://www.microsoft.com/en-us/research/academic-program/microsoft-quantum-challenge/challenge/.

For each task, we have provided Jupyter Notebooks that served as starting points for the participants to learn how to program quantum annealers and to successfully carry out the tasks following the submission guidelines. Moreover, we provided the slides that were presented during the ECIR Tutorial [10] covering the fundamental concepts of QC and QA. We also streamed and recorded a video tutorial[4] about the usage of our infrastructure and the notebooks available to the participants.

For both tasks, participants are asked to submit their runs using both QA and SA. In this way, it will be possible to compare the efficiency and effectiveness of these two similar optimization techniques that employ quantum annealers and traditional hardware respectively.

3.1 Task 1 Quantum Feature Selection

This task focuses on formulating the well-known *NP-Hard* feature selection problem in such a way that it can be solved with a quantum annealer, similarly to what has already been done in previous works [9,18].

Objectives. Feature Selection is a widespread problem for both IR and RS which requires the identification of a subset of the available features (e.g., the most informative, less noisy, etc.) to train a learning model. This problem is very impacting since many of IR and RS systems involve the optimization of learning models, and reducing the dimensionality of the input data can improve their performance. Therefore, in this task, we aim to understand if QA can be applied to solve this problem more efficiently and effectively, exploiting its capability of exploring a larger problem space in a short amount of time.

Sub-tasks. Task 1 is divided into two sub-tasks:

- **Task 1A**: Feature Selection for IR. This task involves selecting the optimal subset of features using QA and SA that will be used to train a LambdaMART [7] model according to a Learning-To-Rank framework;
- **Task 1B**: Feature Selection for RS. This task involves selecting the optimal subset of features using QA and SA that will be used to train a kNN recommendation system model. The item-item similarity is computed with cosine on the feature vectors, a shrinkage of 5 is added to the denominator and the number of selected neighbors for each item is 100.

Datasets. For Task 1A, we decided to employ the famous MQ2007 [23] and the Istella S-LETOR [16] datasets. MQ2007 represents an easier challenge since it has 46 features, allowing direct embedding of the problem formulations inside the QPU of quantum annealers. Istella instead has 220 features and it is impossible

[4] https://www.youtube.com/watch?v=fKrnaJn40Kk/.

to embed problem formulations directly, thus requiring some further processing steps for the participants to fit the problem into the physical QPU hardware.

For Task 1B instead, we decided to employ a custom dataset of music recommendations containing 1.9 thousand users and 18 thousand items. The dataset contains both collaborative data, with 92 thousand implicit user-item interactions, as well as two different sets of item features that are derived from item descriptions and user-provided tags, called Item Content Matrix (ICM). The small set, ICM_150, includes 150 features and can be embedded directly on the QPU with small adjustments, the large set, ICM_500, has 500 features and requires significant pruning to fit in the QPU or the use of Hybrid methods. Both sets of features contain noisy and redundant features.

Evaluation Measures. The official evaluation measure for both Task 1A and Task 1B is nDCG@10.

Baseline. For sub-task 1A the baseline is a Feature Selection model that uses a Recursive Feature Elimination approach paired with a Linear Regression model to select the most relevant subset of features.

For sub-task 1B the baseline is a kNN recommendation system model that uses all the available features. The hyperparameters are the same used for the model computed on the selected features, i.e., the item-item similarity is computed with cosine adding a shrink term of 5 to the denominator, and the number of neighbors is 100.

Runs Format. Participants in both tasks 1A and 1B can submit a maximum of 5 runs per dataset using QA or Hybrid methods and a maximum of 5 runs using SA. Each run that uses QA or Hybrid methods should correspond to a run that employs SA. In this way, it is possible to make a fair comparison between them.

The results of the run must be a text file which lists the features that were selected, one per line. The discarded features are not reported in the run file. Furthermore, the last line must report the list of IDs associated with the problems solved using QA, SA, or Hybrid to obtain the final subset of features by the considered approach.

Each run file must be left in each team's workspace in a specific directory called */config/workspace/submissions*, which is already available.

The submission file name should comply with the format
[Task]_[Dataset]_[Method]_[Groupname]_[SubmissionID].txt, where:

- [**Task**]: it should be either *1A* or *1B* based on the task the submission refers to;
- [**Dataset**]: it should be either *MQ2007*, *Istella*, *150_ICM* or *500_ICM* based on the dataset used;
- [**Method**]: it should be either *QA* or *SA* based on the method used;
- [**Groupname**]: the team name;

- [**SubmissionID**]: a custom submission ID that must be the same for the submissions using the same algorithm but performed with different methods (e.g., QA or SA).

3.2 Task 2 - Quantum Clustering

This task focuses on the formulation of the Clustering problem in such a way that it can be solved with a quantum annealer. It involves grouping the items according to their characteristics. Thus, "similar" items fall into the same group while different items belong to distinct groups.

Objectives. Clustering is a relevant problem for IR and RS since it can be helpful for organizing large collections, helping users explore a collection, and providing similar search results to a given query. Furthermore, it can be beneficial to split users according to their interests or build user models with the cluster centroids [27] speeding up the runtime of the system or its effectiveness for users with limited data.

This task is more focused on the IR field and is applied in a document retrieval scenario where documents have been transformed into their corresponding embeddings by a Transformer model. Each document can be seen as a vector in the space and it is possible to cluster points based on their distances, which can be interpreted as a dissimilarity function: the more distant two vectors are, the more different the corresponding documents are likely to be. In this task, participants should apply QA and SA to cluster documents into 10, 25, and 50 clusters. Participants must report the found centroids and the corresponding associated documents.

By clustering documents, it is possible to reduce the searching time by considering the most similar centroid to the input query and then retrieving only the documents belonging to that centroid's cluster instead of looking at the whole collection of documents.

Clustering fits very well with a QUBO formulation and various methods have already been proposed [3,4,25]. Most of these methods involve the usage of one variable per document, thus making it very hard to consider large datasets due to the limited number of physical qubits and interconnections between them. There are ways to overcome this issue, such as by applying a coarsening or a hierarchical approach.

Datasets. For this task, we considered a custom split of the ANTIQUE [15] dataset containing 6486 documents, 200 queries, and manual relevance judgments. Each document and each query have been transformed into a corresponding embedding with the pre-trained **all-mpnet-base-v2** model[5]. The queries are divided into 50 for the Training Dataset and 150 for the Test Dataset.

[5] https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

Evaluation Measures. The official evaluation measures for Task 2 are:

- the Davies-Bouldin Index to measure the overall cluster quality without considering the document retrieval phase;
- nDCG@10 to measure the retrieval effectiveness based on the clusters found.

Baseline. For this task, the baseline is a traditional k-Medoids approach using the cosine distance as a distance function.

Runs Format. Participants in task 2 can submit a maximum of 5 runs for each number of clusters (i.e., 10, 25, 50) using QA or Hybrid methods and a maximum of 5 runs using SA. Each run that uses QA or Hybrid methods should correspond to a run that employs SA. In this way, it is possible to make a fair comparison between them.

The run file must be a text file (JSON formatted) with a list of 10, 25, and 50 vectors that represent the final centroids achieved through their clustering algorithm. Each centroid should also be followed by the list of documents that belong to the given cluster. Furthermore, the last line must report the list of IDs associated with the problems solved using QA, SA, or Hybrid to obtain the final clusters by the considered approach.

Each run file must be left in each team's workspace in a specific directory called */config/workspace/submissions*, which is already available.

The submission file name should comply with the format *[Centroids]_[Method]_[Groupname]_[SubmissionID].txt*, where:

- [**Centroids**]: it should be either 10, 25, or 50 based on the number of centroids;
- [**Method**]: it should be either *QA* or *SA* based on the method used;
- [**Groupname**]: the team name;
- [**SubmissionID**]: a custom submission ID that must be the same for the submissions using the same algorithm but performed with different methods (e.g., QA or SA).

4 Lab Setup

In this section, we detail the infrastructure that was specifically created to carry out this lab and we present the guidelines the participants had to comply with to submit their runs.

4.1 Infrastructure

Having access to quantum annealers is not straightforward. In fact, D-Wave enforces some policies on the usage of these devices by setting some monthly timing quotas to submit and solve problems on their devices. There are API

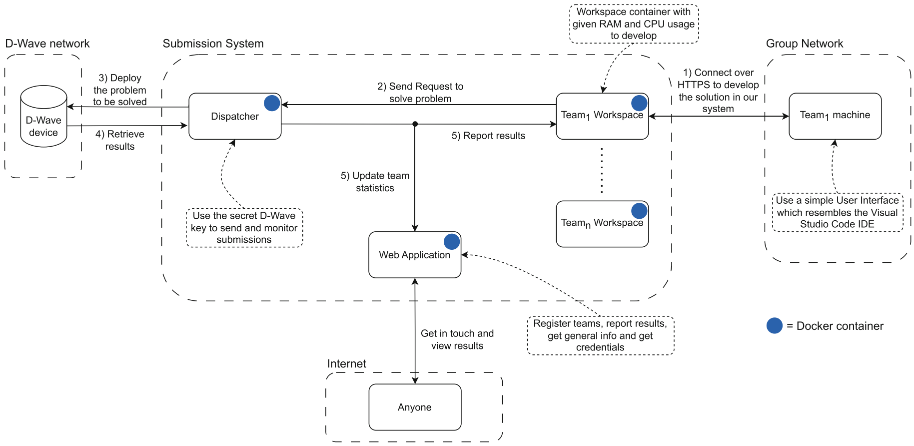


Fig. 1. High-level representation of the infrastructure.

keys that are given to people who use quantum annealers so that it will be possible to monitor the access and usage.

Since it is not possible to disclose our API key to the participants, we decided to build our own infrastructure that allows participants to use quantum annealers without knowing our API key and without needing to stipulate any agreements with D-Wave to obtain their own API keys.

Furthermore, to measure efficiency participants must use the same computing hardware. To this end, our infrastructure provides all the participants with corresponding workspaces located in an AWS server. All workspaces have the same computational resources in terms of CPU and RAM, thus ensuring also easy reproducibility.

Finally, we wanted to create a workflow that was as easy as possible. To this end, participants can access our infrastructure directly from the Web through a simple interface. This interface lets them monitor their quotas but also allows them to develop and execute their code directly from their browsers, without having to worry about installing anything on their machines or dealing with command-line tools.

This infrastructure has been implemented using Docker images orchestrated through Kubernetes. It is made up of several components that are interconnected together to provide both organizers and participants easy access to the needed resources, see Fig. 1. All problems submitted by the participants were saved in a database to monitor their quotas and to gather data to draw statistics about the lab.

The final infrastructure was deployed on a *m6a.8xlarge* AWS EC2 instance equipped with an AMD EPYC 7R13 processor. Table 1 reports the specifications of the hardware resources corresponding to that instance and to each team's workspace. All participants were given the same monthly quota to use quantum resources. Table 2 reports the monthly quotas according to the two tasks.

Table 1. The hardware resources corresponding to the AWS EC2 instance and to the participants' workspaces.

Hardware resources			
–	CPU	RAM	Hard Drive
Infrastructure	32 cores	128 GB RAM	1 TB HDD
Workspace	1200 millicores	10 GB RAM	20 GB HDD

Table 2. The monthly quotas to use quantum resources according to the tasks.

Monthly quotas for the tasks			
Task	March	April	May
Task 1: Feature Selection	30 s	30 s	50 s
Task 2: Clustering	50 s	50 s	150 s

4.2 General Guidelines

Each team has access to its personal area inside our infrastructure with the credentials that have been provided to them. All runs must be executed by using the workspaces that have been created for each one of the participating teams, thus ensuring a fair comparison and easy reproducibility.

All participants cannot exceed their given quotas (see Table 2) to execute problems on quantum devices. The quotas can be monitored by each participating team through a dashboard that is constantly being automatically updated, reporting usages of the different methods (i.e., QA, H, and SA) and some general statistics.

All participants' runs must follow the file formats that are already described in Sect. 3.1 and 3.2 to allow us running our evaluation tools smoothly.

Participants have also been asked to upload their files on their own Bitbucket git repositories to enhance reproducibility. Each repository has been created by us inside a Bitbucket project[6]. Their repositories have been kept private through the challenge but are now public.

5 Results

In this Section, we present the results achieved by the participants and we discuss their approaches. Out of the 26 registered teams, 7 teams managed to upload some final runs. In total, the number of runs is 65 considering both SA, QA, and H(H was introduced in Sect. 2.1). Table 3 reports the 7 teams that correctly participated and submitted some final runs.

In total, throughout the entire lab participants have submitted 976 problems. Specifically, 758 of them were solved with SA, while 199 were solved using QA

[6] https://bitbucket.org/eval-labs/workspace/projects/QCLEF24.

Table 3. The teams who participated and submitted at QuantumCLEF 2024.

Team	Affiliation	Country
BIT.UA	IEETA/DETI, LASI, University of Aveiro	Portugal
CRUISE	RMIT University	Australia
NICA	Iran University of Science and Technology, Departement of Computer Engineering	Iran, Islamic Republic Of
OWS	Friedrich Schiller Universität Jena	Germany
qIIMAS	Universidad Nacional Autonoma de Mexico	Mexico
QTB	Universidad Tecnologica de Bolivar	Colombia
shm2024	Madras Christian College, Chennai	India

and 18 with the H method. The total execution time of SA has been almost 12 h while the total QA and H execution time has been roughly 4 min.

The QA execution time in this whole Section refers to the *Annealing* phase as described in Sect. 2.1, therefore it includes the time required to program the QPU, sampling, and reading the result. The embedding time and network latencies are not taken into account and are left to be considered for possible future editions of the QuantumCLEF lab.

5.1 Task 1A

Here we present the results achieved by the teams participating in task 1A.

MQ2007 Dataset. As it is possible to see in Table 4, teams considered different numbers of features in their submissions. In general, we can observe that most of the submissions achieve similar nDCG@10 values when considering a number of features that lies between 10 and 25. In fact, Fig. 2 shows that for these runs the Tukey HSD test performed after the Two-Way ANOVA hypothesis test shows no significant differences. Instead, runs that consider only 5 features achieve nDCG@10 values that are significantly different (lower) with respect to the others. This is reasonable since by considering too few features, then there is a high information loss.

Figure 3 shows the nDCG@10 values and Annealing timings of the runs that used QA and SA. From this figure we can see that, in terms of efficiency (i.e., Annealing time), runs using QA required a shorter amount of time with respect to SA. On average, QA required ≈ 9.89 times less compared to SA, thus representing a more efficient alternative. Considering effectiveness, SA seems to be performing more consistently. However, on average it performs only ≈ 1.03 times better compared to QA.
Teams adopted different approaches to address this task:

- team **BIT.UA** [1] tried different QUBO formulations that involved the usage of different correlation-based measures such as Spearman coefficient, Pearson coefficient, and Mutual Information [9]. Furthermore, their approach also

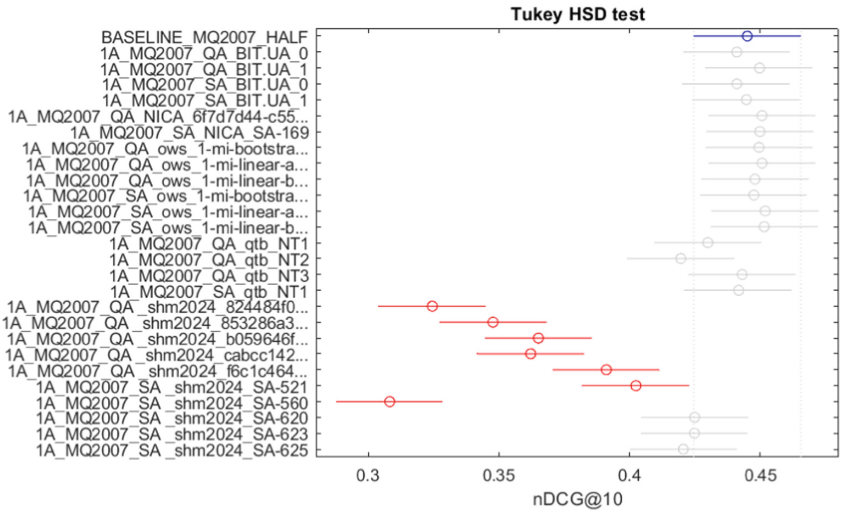
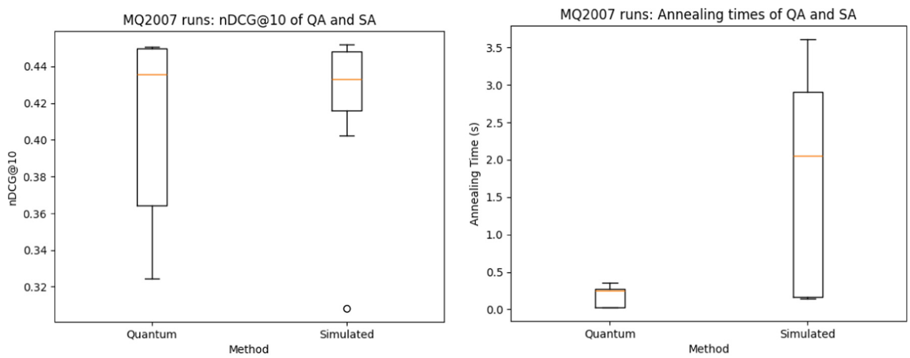


Fig. 2. The Tukey HSD test considering the nDCG@10 values associated with different runs and queries for the MQ2007 dataset.

involved the usage of a scaling factor to automatically balance the importance of the diagonal terms in the matrix Q with respect to the off-diagonal terms. Additionally, they also tried investigating some non-linear functions that adjusted the weights of the values returned by the correlation-based measures. The number of features chosen was decided by using a validation dataset approach with a custom LambdaMART model.

- team **NICA** [17] and team **shm2024** [13] used a QUBO formulation which involved the Mutual Information [9] as a correlation-based measure.
- team **QTB** [22] investigated different QUBO formulations involving different correlation-based measures (e.g., Mutual Information [9]). The team employed all methods (i.e., QA, H and SA), and the H approach allowed them to achieve a high score with only a few features thanks to its pre-processing and post-processing capabilities.
- team **OWS** [12] employed a QUBO matrix that was formulated using Mutual Information [9], in which some of its components were recalculated using the results achieved by a bootstrapping approach. In this way, the team recalculated the values associated with the diagonal components, the off-diagonal components, or both. The team focused on choosing only 25 features and the optimization of the number of considered features is left for future works.

Istella Dataset. As it is possible to see in Table 5 and in Fig. 4, also in this case teams considered different numbers of features in their submissions. However, for the Istella dataset, most of the runs are statistically different from each other because the number of features used varies a lot. It is interesting to see that the



(a) nDCG@10 of SA and QA runs (b) Annealing time of SA and QA runs

Fig. 3. The box plots of the nDCG@10 values and Annealing timings associated with the runs using QA and SA on the MQ2007 dataset.

baseline method employing Recursive Feature Elimination considering 110 features performed much worse with respect to all participants' runs. Furthermore, running Recursive Feature Elimination to keep the top 110 features required a considerable amount of time (almost 2 h of computation) and a considerable amount of RAM (24 GB), which is much higher than the teams' workspace specifications.

The teams adopted similar approaches to the ones described for the MQ2007 dataset to solve the Feature Selection task on the Istella dataset. However, since the dataset could not fit entirely in the QPU due to the high number of features, two teams decided to adopt the following pre-processing techniques:

- team **BIT.UA** [1] employed different approaches such as using a first stage SA approach to select only a subset of features or the manual elimination of features with high correlation values between them before solving the problem with QA.
- team **NICA** [17] kept only the 50 features that had the highest Mutual Information value towards the target variable, thus reducing the feature set.

Figure 5 shows the nDCG@10 values and Annealing timings of the runs that used QA and SA. From this figure we can see that, in terms of efficiency (i.e., Annealing time), also in this case runs using QA required a shorter amount of time with respect to SA. On average, QA required ≈ 10.45 times less compared to SA, thus representing a more efficient alternative. Similar considerations apply also for effectiveness. In fact, SA seems to be performing more consistently however, on average it performs only ≈ 1.03 times better compared to QA.

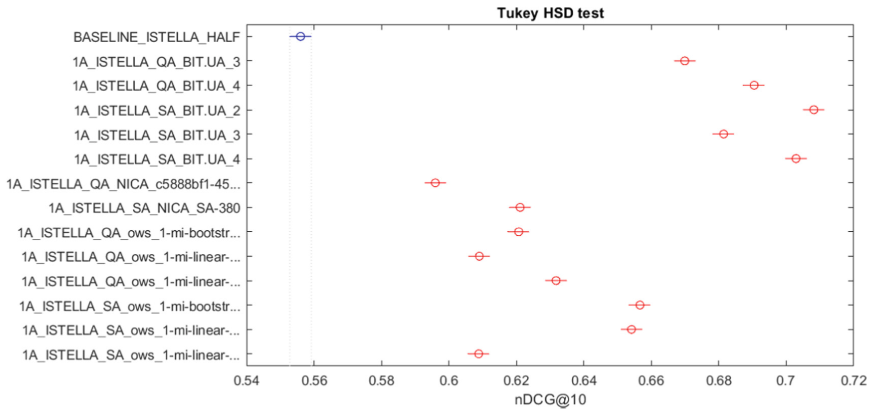
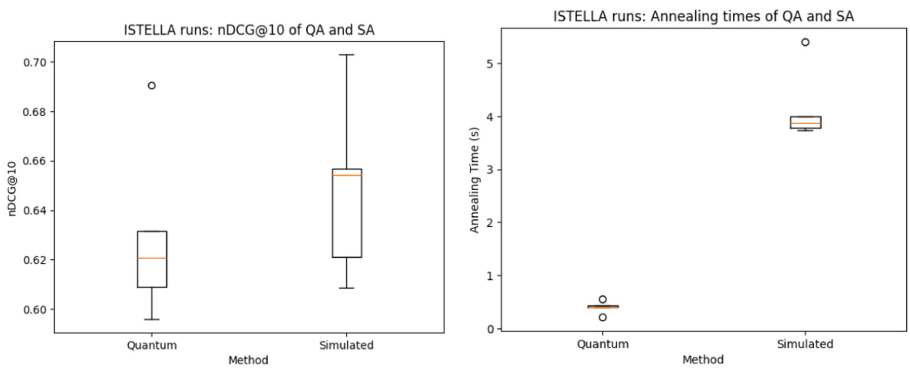


Fig. 4. The Tukey HSD test considering the nDCG@10 values associated with different runs and queries for the Istella dataset.

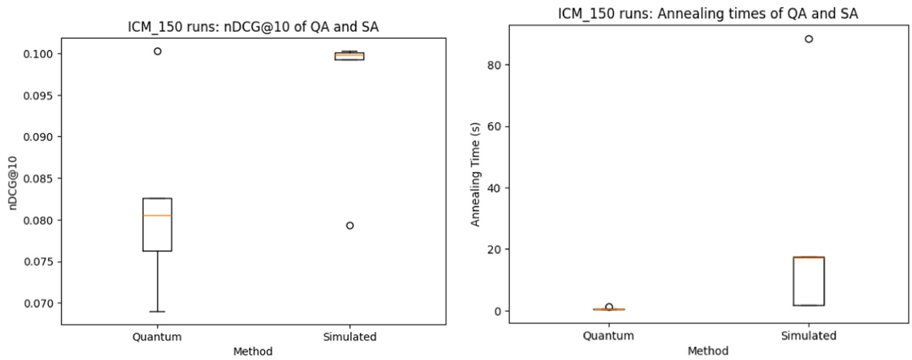


(a) nDCG@10 of SA and QA runs (b) Annealing time of SA and QA runs

Fig. 5. The box plots of the nDCG@10 values and Annealing timings associated with the runs using QA and SA on the Istella dataset.

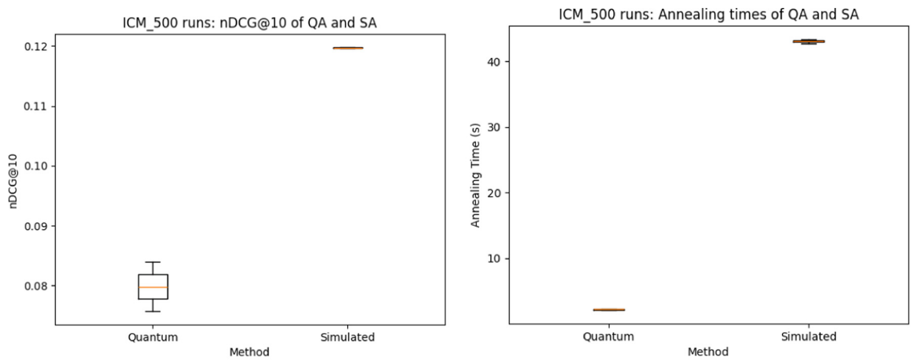
5.2 Task 1B

Here we present the results achieved by the two teams participating in task 1B. Results are divided according to the two feature sets. For both the small ICM (see Table 6) and the large one (see Table 7) the teams were able to improve the effectiveness of the baseline RS by a large margin, around 23% on the small set and 44% on the large one. Team **CRUISE** [19] especially achieved a large improvement by developing a counterfactual version of nDCG to enhance a feature selection method based on Mutual Information. The idea considers that Mutual Information does not account for the final goal of making recommendations.



(a) nDCG@10 of QA and SA runs (b) Annealing time of QA and SA runs

Fig. 6. The box plots of the nDCG@10 values and Annealing timings associated with the runs using QA and SA on the ICM_150 dataset.



(a) nDCG@10 of QA and SA runs (b) Annealing time of QA and SA runs

Fig. 7. The box plots of the nDCG@10 values and Annealing timings associated with the runs using QA and SA on the ICM_500 dataset.

The proposed approach is based on MIQUBO [9] and introduces a term in the diagonal of Q which represents the change in nDCG@10 obtained by removing each of the features individually, weighted by a scaling factor. In this way, the diagonal of Q includes both the Mutual Information between the feature values and the target label, as well as the weighted change in nDCG@10. For the small ICM, with 150 features, QA is 35.88 times faster than SA but it is 1.17 times worse in terms of nDCG@10 (see Fig. 6).

For the large ICM, with 500 features, that could not fit on the QPU, team **CRUISE** [19] split the features into subsets small enough to be tackled by the QPU. Then, the features selected in each subset have been merged into a final set of features. For this large ICM QA is 19.53 times faster than SA but it is 1.5 times worse in terms of nDCG@10 (see Fig. 7). Note that the number of selected features is very different so this could play a role.

5.3 Task 2

Here we present the results achieved by the teams participating in task 2. Table 8 reports the results achieved in this task.

In this task, we can see that team **qIIMAS** managed to achieve higher results with respect to the baseline for each number of clusters considered. The approach adopted by team **qIIMAS** [2] consisted of employing the QUBO formulation proposed in a previous work [5]. Due to the high dimensionality of the dataset, they decided to first apply a traditional approach to reduce the number of points n to some representatives m where $m < n$. Then they performed the clustering approach on the m representatives in a hierarchical fashion, returning the final set of centroids and their associated n points. They investigated the usage of both QA, H, and SA.

In Fig. 8 we can observe that there are no statistical differences among runs using H and runs using SA considering the nDCG@10 values achieved.

Figure 9 shows the Annealing time of the runs that used H and SA. From this figure we can see that, in terms of efficiency (i.e., Annealing time), runs using H required a shorter amount of time with respect to SA. On average, H required ≈ 21.75 times less compared to SA, thus representing a more efficient alternative. In addition, the H methods achieved slightly better results in terms of effectiveness, being ≈ 1.02 times better than SA on average.

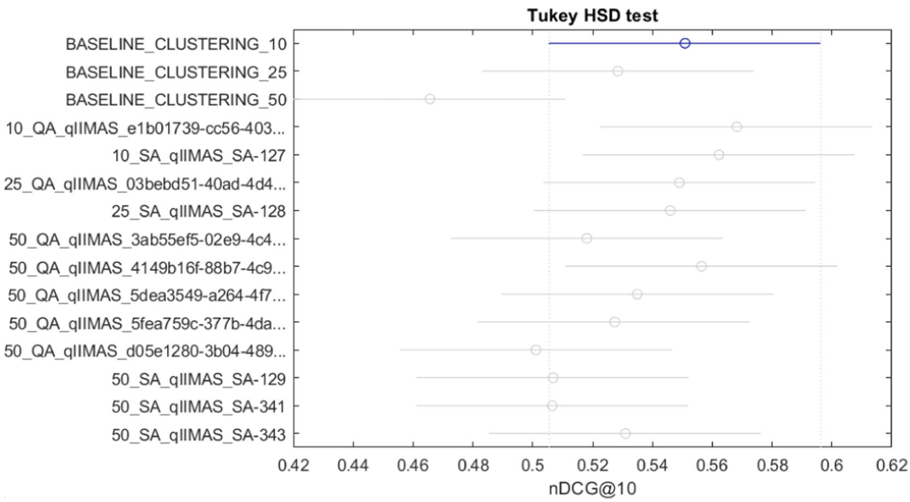
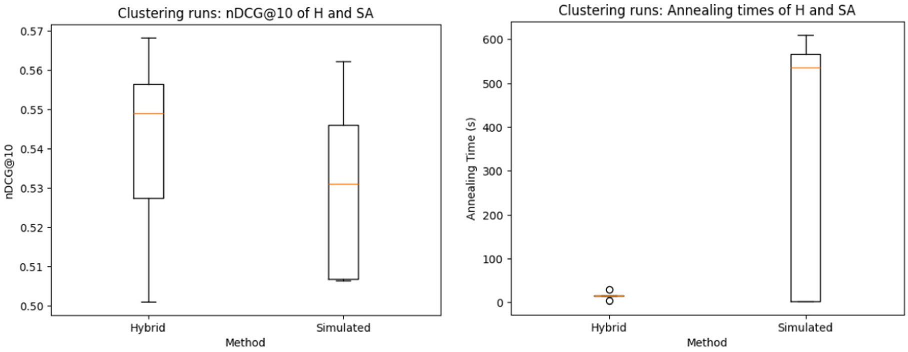


Fig. 8. The Tukey HSD test considering the nDCG@10 values associated with different runs and queries for the Clustering dataset.



(a) nDCG@10 of H and SA runs (b) Annealing time of H and SA runs

Fig. 9. The box plots of the nDCG@10 values and Annealing timings associated with the runs using H and SA on the Clustering dataset.

6 Conclusions and Future Work

In this paper, we have presented the overview of the first edition of the QuantumCLEF 2024 lab, the first lab at CLEF focusing on the study, development, and evaluation of QC algorithms.

This lab was composed of two tasks concerning the problems of Feature Selection and Clustering, specifically focused on IR and RS systems. An ad-hoc infrastructure was created to ease the participants' workflow and to grant them access to computational resources and the cutting-edge quantum annealers provided by D-Wave.

A total of 26 teams registered for the lab and 7 of them successfully managed to submit their runs. The results have shown that QA and H managed to achieve comparable results in terms of effectiveness with respect to SA while achieving a higher level of efficiency in terms of Annealing time. This shows that QC is starting to become a powerful technology that could help in the resolution of complex problems, especially in the future once it has matured enough.

This lab represented a great opportunity not only to develop and evaluate QC algorithms on **real quantum computers** (quantum technologies are still not easily accessible to the general public) but also to raise awareness of the potential of QC, which is likely to become a powerful technology in the future. The data obtained throughout the challenge has also been useful in preparing a new QC tutorial presented to the community at the international SIGIR conference 2024 [11]. Furthermore, participants were provided with comprehensive materials such as videos, slides, and examples that allowed them to learn how QC and QA work. Finally, we opted for maximum transparency, allowing participants to work with the actual D-Wave libraries without constraining them to use custom functions. In this way, participants familiarized themselves with the official D-Wave libraries and, thus, are now able to program quantum annealers even outside our infrastructure to solve other problems in their research field.

In the future, we plan to organize a second edition of QuantumCLEF with different tasks and more challenges. We also plan to further improve the infrastructure according to the comments received by the participants through the lab to ensure a smoother experience for participants of a possible future edition of QuantumCLEF. Moreover, we would like to invest in a more powerful infrastructure that will grant access to more participants and that will provide more resources (in terms of CPU and RAM) to each workspace. In this way, it will be possible to consider even a more fair comparison between SA and QA. If possible, we would also like to extend the infrastructure to include a gate-based quantum computer [24], in addition to the already available quantum annealer.

Acknowledgments. We acknowledge the financial support from ICSC - "National Research Centre in High Performance Computing, Big Data and Quantum Computing", funded by the European Union - NextGenerationEU.

We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support.

A Task 1A - Team Results

Table 4. The results for Task 1A on the MQ2007 dataset. An adjacent couple of rows (marked with the same color) represents the results achieved with QA/H and SA using the same problem formulation. Results marked in yellow(●) refer to the baselines' results.

Team	Submission id	nDCG@10	Annealing time (ms)	Type	N features
BIT.UA	1A_MQ2007_QA_BIT.UA_0	0.441	274	QA	18
BIT.UA	1A_MQ2007_SA_BIT.UA_0	0.441	1351	SA	16
BIT.UA	1A_MQ2007_QA_BIT.UA_1	0.4497	270	QA	20
BIT.UA	1A_MQ2007_SA_BIT.UA_1	0.4446	3607	SA	18
NICA	1A_MQ2007_QA_NICA_6f7d7d44-c559-4e36-9b10-b7e51e521036	0.4506	274	QA	17
NICA	1A_MQ2007_SA_NICA_SA-169	0.4498	3510	SA	15
OWS	1A_MQ2007_QA_ows_1-mi-bootstrap-mixture	0.4495	279	QA	25
OWS	1A_MQ2007_SA_ows_1-mi-bootstrap-mixture	0.4475	2818	SA	25
OWS	1A_MQ2007_QA_ows_1-mi-linear-and-quadratic-bootstrapped-boost-3	0.4506	270	QA	25
OWS	1A_MQ2007_SA_ows_1-mi-linear-and-quadratic-bootstrapped-boost-3	0.4519	2752	SA	25
OWS	1A_MQ2007_QA_ows_1-mi-linear-bootstrapped-boost-3	0.448	241	QA	25
OWS	1A_MQ2007_SA_ows_1-mi-linear-bootstrapped-boost-3	0.4515	2759	SA	25
QTB	1A_MQ2007_QA_qtb_NT1	0.4299	356	QA	13
QTB	1A_MQ2007_SA_qtb_NT1	0.4024	3174	SA	10
QTB	1A_MQ2007_QA_qtb_NT2	0.4195	5000	H	10
QTB	-	-	-	SA	-
QTB	1A_MQ2007_QA_qtb_NT3	0.443	4309	H	10
QTB	-	-	-	SA	-
shm2024	1A_MQ2007_QA _shm2024_b059646f-a9fd-4fd6-9589-c6e117400a9e	0.365	30	QA	5
shm2024	1A_MQ2007_SA _shm2024_SA-521	0.4024	284	SA	5
shm2024	1A_MQ2007_QA _shm2024_cabcc142-3fc5-4b22-8a6b-c7a45857fbc2	0.3621	27	QA	5
shm2024	1A_MQ2007_SA _shm2024_SA-560	0.3082	164	SA	5
shm2024	1A_MQ2007_QA _shm2024_f6c1c464-6dba-4a44-93b8-92ad6c4f60f9	0.391	29	QA	5
shm2024	1A_MQ2007_SA _shm2024_SA-620	0.4249	143	SA	5
shm2024	1A_MQ2007_QA _shm2024_853286a3-7f47-4de8-b0a0-247a65e6f6b6	0.3477	28	QA	5
shm2024	1A_MQ2007_SA _shm2024_SA-623	0.4248	147	SA	5
shm2024	1A_MQ2007_QA _shm2024_824484f0-b6fa-44b6-9bc7-0cb073db84e7	0.3245	29	QA	5
shm2024	1A_MQ2007_SA _shm2024_SA-625	0.4205	144	SA	5
BASELINE	ALL_FEATURES	0.4473	-	-	46
BASELINE	RFE HALF_FEATURES	0.4450	-	-	23

Table 5. The results for Task 1A on the Istella dataset. An adjacent couple of rows (marked with the same color) represents the results achieved with QA/H and SA using the same problem formulation. Results marked in yellow(●) refer to the baselines' results.

Team	Submission id	nDCG@10	Annealing time (ms)	Type	N features
BIT.UA	1A_Istella_QA_BIT.UA_3	0.6699	16325	SA+QA	92
BIT.UA	1A_Istella_SA_BIT.UA_3	0.6814	19071	SA	90
BIT.UA	1A_Istella_QA_BIT.UA_4	0.6905	551	QA	82
BIT.UA	1A_Istella_SA_BIT.UA_4	0.7029	5404	SA	72
BIT.UA	-	-	-	SA	-
BIT.UA	1A_Istella_SA_BIT.UA_2	0.7081	13827	SA	161
NICA	1A_Istella_QA_NICA_c5888bf1-4549-418c-92b8-b7175c9185e4	0.596	427	QA	15
NICA	1A_Istella_SA_NICA_SA-380	0.6211	3998	SA	15
OWS	1A_Istella_QA_ows_1-mi-bootstrap-mixture	0.6207	215	QA	25
OWS	1A_Istella_SA_ows_1-mi-bootstrap-mixture	0.6566	3875	SA	25
OWS	1A_Istella_QA_ows_1-mi-linear-and-quadratic-bootstrapped-boost-3	0.609	394	QA	25
OWS	1A_Istella_SA_ows_1-mi-linear-and-quadratic-bootstrapped-boost-3	0.6541	3728	SA	25
OWS	1A_Istella_QA_ows_1-mi-linear-bootstrapped-boost-3	0.6317	402	QA	25
OWS	1A_Istella_SA_ows_1-mi-linear-bootstrapped-boost-3	0.6088	3785	SA	25
BASELINE	ALL_FEATURES	0.7146	-	-	220
BASELINE	RFE HALF_FEATURES	0.5560	-	-	110

B Task 1B - Team Results

Table 6. Task 1B results on the 150_ICM dataset. Adjacent row pairs (same color) show the results achieved with QA/H and SA for the same problem formulation. Results highlighted in yellow(●) refer to the baselines' results.

Team	Submission id	nDCG@10	Annealing time (ms)	Type	N features
CRUISE	1B_150_ICM_QA_CRUISE_1	0.0805	536	QA	138
CRUISE	1B_150_ICM_SA_CRUISE_1	0.0998	1745	SA	140
CRUISE	1B_150_ICM_QA_CRUISE_2	0.0826	529	QA	136
CRUISE	1B_150_ICM_SA_CRUISE_2	0.0993	17358	SA	140
CRUISE	1B_150_ICM_QA_CRUISE_3	0.0690	531	QA	132
CRUISE	1B_150_ICM_SA_CRUISE_3	0.1001	1760	SA	140
CRUISE	1B_150_ICM_QA_CRUISE_4	0.0763	558	QA	133
CRUISE	1B_150_ICM_SA_CRUISE_4	0.0793	17387	SA	140
CRUISE	1B_150_ICM_QA_CRUISE_5	0.1003	1375	QA	144
CRUISE	1B_150_ICM_SA_CRUISE_5	0.1003	88395	SA	144
NICA	-	-	-	QA	-
NICA	1B_150_ICM_SA_NICA_SA-457	0.0895	12247	SA	145
BASELINE	ALL_FEATURES	0.0810	-	-	150

Table 7. Task 1B results on the 500_ICM dataset. Adjacent row pairs (same color) show the results achieved with QA/H and SA for the same problem formulation. Results highlighted in yellow(●) refer to the baselines' results.

Team	Submission id	nDCG@10	Annealing time (ms)	Type	N features
CRUISE	1B_500_ICM_QA_CRUISE_1	0.0757	2287	QA	407
CRUISE	1B_500_ICM_SA_CRUISE_1	0.1196	43339	SA	450
CRUISE	1B_500_ICM_QA_CRUISE_2	0.0839	2123	QA	397
CRUISE	1B_500_ICM_SA_CRUISE_2	0.1198	42777	SA	450
BASELINE	ALL_FEATURES	0.0827	-	-	500

C Task 2 - Team Results

Table 8. Task 2 results. Adjacent row pairs (same color) show the results achieved with QA/H and SA for the same problem formulation. Results highlighted in yellow(●) refer to the baselines' results.

Team	Submission id	nDCG@10	Annealing time (ms)	Type	N features
CRUISE	1B_500_ICM_QA_CRUISE_1	0.0757	2287	QA	407
CRUISE	1B_500_ICM_SA_CRUISE_1	0.1196	43339	SA	450
CRUISE	1B_500_ICM_QA_CRUISE_2	0.0839	2123	QA	397
CRUISE	1B_500_ICM_SA_CRUISE_2	0.1198	42777	SA	450
BASELINE	ALL_FEATURES	0.0827	-	-	500

References

1. Almeida, T., Matos, S.: Towards a hyperparameter-free QUBO formulation for feature selection in IR. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
2. Alvarez Giron, W., Tellez, J., Tovar Cortes, J., Gómez Adorno, H.: Team qIIMAS on task 2 - clustering. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
3. Arthur, D., Date, P.: Balanced k-means clustering on an adiabatic quantum computer. Quantum Inf. Process. **20**(9), 294 (2021). https://doi.org/10.1007/s11128-021-03240-8
4. Bauckhage, C., Piatkowski, N., Sifa, R., Hecker, D., Wrobel, S.: A QUBO formulation of the k-medoids problem. In: Lernen, Wissen, Daten, Analysen, Berlin, Germany. CEUR Workshop Proceedings, vol. 2454, pp. 54–63. CEUR-WS.org (2019). https://ceur-ws.org/Vol-2454/paper_39.pdf
5. Bauckhage, C., Piatkowski, N., Sifa, R., Hecker, D., Wrobel, S.: A QUBO formulation of the k-medoids problem. In: LWDA, pp. 54–63 (2019)

6. Bertsimas, D., Tsitsiklis, J.: Simulated annealing. Stat. Sci. **8**(1), 10–15 (1993)
7. Burges, C.J.C.: From RankNet to LambdaRank to LambdaMART: an overview. Technical report, Microsoft Research, MSR-TR-2010-82 (2010)
8. Cai, J., Macready, W.G., Roy, A.: A practical heuristic for finding graph minors. arXiv preprint arXiv:1406.2741 (2014)
9. Ferrari Dacrema, M., Moroni, F., Nembrini, R., Ferro, N., Faggioli, G., Cremonesi, P.: Towards feature selection for ranking and classification exploiting quantum annealers. In: Proceedings of the 45th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2022), pp. 2814–2824. ACM Press, New York (2022)
10. Ferrari Dacrema, M., Pasin, A., Cremonesi, P., Ferro, N.: Quantum computing for information retrieval and recommender systems. In: Goharian, N., et al. (eds.) ECIR 2024. LNCS, vol. 14612, pp. 358–362. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-56069-9_47
11. Ferrari Dacrema, M., Pasin, A., Cremonesi, P., Ferro, N.: Using and evaluating quantum computing for information retrieval and recommender systems (2024)
12. Fröbe, M., Alexander, D., Hendriksen, G., Schlatt, F., Hagen, M., Potthast, M.: Team OpenWebSearch at CLEF 2024: QuantumCLEF. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
13. Gersome, S., Mahibha, J., Thenmozhi, D.: Team shm2024 on quantum feature selection. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
14. Glover, F., Kochenberger, G., Hennig, R., Du, Y.: Quantum bridge analytics I: a tutorial on formulating and using QUBO models. Ann. Oper. Res. **314**, 141–183 (2022)
15. Hashemi, H., Aliannejadi, M., Zamani, H., Croft, W.B.: ANTIQUE: a non-factoid question answering benchmark. In: Jose, J.M., et al. (eds.) ECIR 2020, Part II. LNCS, vol. 12036, pp. 166–173. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45442-5_21
16. Lucchese, C., Nardini, F.M., Orlando, S., Perego, R., Silvestri, F., Trani, S.: Post-learning optimization of tree ensembles for efficient ranking. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 949–952 (2016)
17. Naebzadeh, A., Eetemadi, S.: Nica at quantum computing CLEF tasks 2024. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
18. Nembrini, R., Ferrari Dacrema, M., Cremonesi, P.: Feature selection for recommender systems with quantum computing. Entropy **23**(8), 970 (2021)
19. Niu, J., Li, J., Deng, K., Ren, Y.: Cruise on quantum computing for feature selection in recommender systems. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
20. Pasin, A., Ferrari Dacrema, M., Cremonesi, P., Ferro, N.: qCLEF: a proposal to evaluate quantum annealing for information retrieval and recommender systems. In: Arampatzis, A., et al. (eds.) CLEF 2023. LNCS, vol. 14163, pp. 97–108. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-42448-9_9

21. Pasin, A., Ferrari Dacrema, M., Cremonesi, P., Ferro, N.: QuantumCLEF - quantum computing at CLEF. In: Goharian, N., et al. (eds.) ECIR 2024. LNCS, vol. 14612, pp. 482–489. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-56069-9_66
22. Payares, E., Puertas, E., Martinez Santos, J.C.: Team QTB on feature selection via quantum annealing and hybrid models. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
23. Qin, T., Liu, T.Y.: Introducing LETOR 4.0 datasets. arXiv preprint arXiv:1306.2597 (2013)
24. Rieffel, E., Polak, W.: An introduction to quantum computing for non-physicists. ACM Comput. Surv. (CSUR) **32**(3), 300–335 (2000)
25. Ushijima-Mwesigwa, H., Negre, C.F.A., Mniszewski, S.M.: Graph partitioning using quantum annealing on the d-wave system. CoRR abs/1705.03082 (2017). http://arxiv.org/abs/1705.03082
26. Van Laarhoven, P.J., Aarts, E.H.: Simulated Annealing. Springer, Dordrecht (1987)
27. Wu, Y., Cao, Q., Shen, H., Tao, S., Cheng, X.: INMO: a model-agnostic and scalable module for inductive collaborative filtering. In: SIGIR 2022: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, pp. 91–101. ACM (2022). https://doi.org/10.1145/3477495.3532000
28. Yarkoni, S., Raponi, E., Bäck, T., Schmitt, S.: Quantum annealing for industry applications: introduction and review. Rep. Prog. Phys. **85**(10), 104001 (2022)

Overview of the CLEF 2024 SimpleText Track

Improving Access to Scientific Texts for Everyone

Liana Ermakova[1], Eric SanJuan[2], Stéphane Huet[2], Hosein Azarbonyad[3], Giorgio Maria Di Nunzio[4], Federica Vezzani[4], Jennifer D'Souza[5], and Jaap Kamps[6(✉)]

[1] Université de Bretagne Occidentale, HCTI, Brest, France
liana.ermakova@univ-brest.fr
[2] Avignon Université, LIA, Avignon, France
[3] Elsevier, Amsterdam, The Netherlands
[4] University of Padua, Padua, Italy
[5] TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
[6] University of Amsterdam, Amsterdam, The Netherlands
kamps@uva.nl
https://simpletext-project.com

Abstract. Everyone acknowledges the importance of objective scientific information. However, finding and understanding relevant scientific documents is often challenging due to complex terminology and readers' lack of prior knowledge. The question is can we improve accessibility for everyone? This paper presents an overview of the SimpleText Track at CLEF 2024 addressing the technical and evaluation challenges associated with making scientific information accessible to a wide audience, including students and non-experts. It describes the data and benchmarks provided for scientific text summarization and simplification, along with the participants' results. The CLEF 2024 SimpleText track is based on four interrelated tasks: Task 1 on *Content Selection*: Retrieving Passages to Include in a Simplified Summary. Task 2 on *Complexity Spotting*: Identifying and Explaining Difficult Concepts. Task 3 on *Text Simplification*: Simplify Scientific Text. Task 4 on *SOTA?*: Tracking the State-of-the-Art in Scholarly Publications.

Keywords: Scientific text simplification · Information extraction · Information retrieval · Natural language processing

1 Introduction

The importance of objective scientific information is universally acknowledged. In practice, accessing, processing, and comprehending relevant scientific documents is challenging due to complex terminology and the potential lack of prior knowledge among readers. The CLEF 2024 SimpleText track aims at improving accessibility to scientific information for everyone, both in terms of information retrieval and natural language processing. The workshop at CLEF 2021 [12] and tracks at CLEF 2022 [16] and 2023 [15] resulted in research community and test collections for improving access to

L. Goeuriot et al. (Eds.): CLEF 2024, LNCS 14959, pp. 283–307, 2024.
https://doi.org/10.1007/978-3-031-71908-0_13

scientific information for everyone. Specifically, test collections for retrieving relevant (and accessible) scientific text [34], for simplifying the language used in scientific documents without compromising the accuracy of the information [13], and for making complex concepts more understandable to a broader audience [11].

Scientific Text Simplification is different from traditional text simplification approaches focusing on lower literacy levels, for example making general text accessible to young readers. Recent advances in IR and NLP hold the promise of removing some of the barriers to scientific information access.[1] The overall impact of CLEF SimpleText is to increase science literacy and broaden the audience of objective, scientific information.

The track's setup is based on the following pipeline: i) select the information to be included in a simplified summary; ii) improve the readability of the scientific text; iii) provide additional background knowledge for remaining difficult concepts; and iv) aggregate information from multiple articles. This results in the following four tasks [16]:

– ***Task 1: Content Selection*** *Retrieving Passages to Include in a Simplified Summary.*
– ***Task 2: Complexity Spotting*** *Identifying and Explaining Difficult Concepts.*
– ***Task 3: Text Simplification*** *Simplify Scientific Text.*
– ***Task 4: SOTA?*** *Tracking the State-of-the-Art in Scholarly Publications.*

A total of 45 teams registered for our SimpleText track at CLEF 2024. A total of 20 teams submitted 207 runs in total. The statistics for these runs submitted are presented in Table 1. However, some runs had problems that we could not resolve. We do not detail them in the paper as well as the 0-scored runs.

This paper gives an overview of the CLEF 2024 SimpleText Track. Further detail per track is provided in the Track overview papers published in the CEUR CLEF Working Notes, specifically for Task 1 on *Content Selection* [35]; Task 2 on *Complexity Spotting* [31]; Task 3 on *Text Simplification* [17]; and Task 4 on *SOTA?* [9].

In the rest of this paper, we will provide a detailed description of each task of the CLEF 2024 edition in four self-contained sections: Task 1: Content Selection in Sect. 2, Task 2: Complexity Spotting in Sect. 3, Task 3: Text Simplification in Sect. 4, and Task 4: SOTA? in Sect. 5. We end with a discussion and conclusions in Sect. 6.

2 Task 1: Retrieving Passages to Include in a Simplified Summary

This section details *Task 1: Content Selection* on Retrieving Passages to Include in a Simplified Summary.

2.1 Description

Given a popular science article targeted to a general audience, this task aims at retrieving passages, which can help to understand this article, from a large corpus of academic abstracts and bibliographic metadata. Relevant passages should relate to any of the topics in the source article.

[1] A joined effort with others like Scholarly Document Processing https://sdproc.org/2024/.

Table 1. CLEF 2024 Simpletext official run submission statistics

Team	Task 1	Task 2			Task 3		Task 4		Total runs
		2.1	2.2	2.3	3.1	3.2	4.1	4.2	
AIIRLab	5	3	3		4	4			19
AMATU							3	9	12
Arampatzis	9	5	5	2	4	4			29
Elsevier	10				8	2			20
L3S							12	12	24
LIA	5								5
PiTheory					11	10			21
Sharigans	1	1	1		1	1			5
SINAI		3	3						6
SONAR					1				1
AB/DPV	1	1	1		1				4
Dajana/Katya		1			1				2
Frane/Andrea		1	1		1				3
Petra/Regina	1	1			1				3
Ruby	1	1			1	1			4
Tomislav/Rowan	2	2			1	1			6
UAmsterdam	6	1		2	4	6			19
UBO	1	1	1		2	2			7
UniPD		3	3						6
UZHPandas					11				11
Total runs	42	24	18	4	52	31	15	21	207

Data. We use popular science articles as a source for the types of topics the general public is interested in and as a validation of the reading level that is suitable for them. The main corpus is a large set of scientific abstracts plus associated metadata covering the fields of computer science and engineering. We reuse the collection of academic abstracts from the Citation Network Dataset (12th version released in 2020)[2] [37]. This collection was extracted from DBLP, ACM, MAG (Microsoft Academic Graph), and other sources. It includes 4,232,520 abstracts in English, published before 2020.

Search requests are based on popular press articles targeted to a general audience, based on *The Guardian* and *Tech Xplore*. Each of these popular science articles represents a general topic that has to be analyzed to retrieve relevant scientific information from the corpus.

We provide the URLs to original articles, the title, and the textual content of each popular science article as a general topic. Each general topic was also enriched with one or more specific keyword queries manually extracted from their content, creating a familiar information retrieval task ranking passages or abstracts in response to a query. Available training data from 2023 includes 29 (train) and 34 (test) queries, with the later set having an extensive recall base due to the large number of submissions in 2023 [34].

[2] https://www.aminer.cn/citation.

In 2024, we introduced two novelties. First, in addition to the ElasticSearch we provided to participants, we made available a new vector database with sentence embedding scores. This database stores for each article their ID, two sentence-embedding vectors computed by all-MiniLM-L6-v2[3] from their title and their abstract. Second, we added between 2 and 5 new queries (with IDs of the form G*.C*) for each of the 20 articles from the Guardian. These topics were generated by ChatGPT 4, with a prompt asking to list the main subtopics related to computer science; they were manually inspected to check they are linked to the original article and are not redundant. They are longer, containing around ten words and focusing on a specific point related to the article. An example of a keyword query is "*system on chip*" (T06.1) and an example of a long query is "*How AI systems, especially virtual assistants, can perpetuate gender stereotypes?*" (G01.C1).

Evaluation. Topical relevance was evaluated with a 0–2 score on the relevance degree towards the content of the original article. We dramatically extended the training qrels with 9,990 additional evaluated documents, with a main focus on the new long queries for the Guardian, and the T06-T11 queries for which no list of relevant documents had yet been provided.

In addition to topical relevance, we took into account other key aspects of the track, such as the text complexity and the credibility of the retrieved results. These evaluations were performed using automatic metrics.

2.2 Participant's Approaches

A total of 11 teams submitted 42 runs in total.

AB/DPV. Varadi and Bartulović [40] submitted 1 run for Task 1. They used our ElasticSearch API and took into account an FKGL readability score for their combined score.

Sharingans. Ali et al. [2] also submitted 1 run. They experimented with the ColBERT neural ranker and used GPT 3.5 to select the most informative and concise passages for inclusion in the summary.

Tomislav/Rowan. Mann and Mikulandric [27] submitted a total of 2 runs. They took the top 100 results retrieved by ElasticSearch. Then, they used cosine similarity on TF-IDF vectors as the relevance score and FKGL score as the combined score.

Petra/Regina. Elagina and Vučić [10] submitted 1 run, for the first 3 queries only, with the same approach as the previous system.

AIIRLab. Largey et al. [25] submitted a total of 5 runs and proposed several models. First, since input queries are short keyword terms, they used query expansion with LLaMA 3 and reranked the top 5,000 results retrieved by TF-IDF with a bi-encoder or a cross-encoder. Second, they applied LLaMa3 as a pairwise re-ranker. Third, they leveraged ElasticSearch with fine-tuned cross-encoders.

UBO. Vendeville et al. [41] submitted a total of 1 run. They used PyTerrier[4] to retrieve documents from TF-IDF scores. Then, the MonoT5 reranker provided by PyTerrier was employed to reorder all extracted documents.

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

[4] https://pyterrier.readthedocs.io/.

UAmsterdam. Bakker et al. [4] submitted a total of 6 runs for Task 1. First, they focused on regular information retrieval effectiveness with 2 vanilla baseline runs on an Anserini index, using either BM25 or BM25+RM3, and 2 other runs generated with neural cross-encoder rerankings of these runs by an MS MARCO-trained ranker. Second, 2 further runs filter out the most complex abstracts per request, using the median FKGL readability measure.

Elsevier. Capari et al. [5] submitted a total of 10 runs. Their approaches mainly centered on creating a ranking model. They started by assessing the performance of several models on a proprietary test collection of scientific papers. Then, the top-performing model was fine-tuned on a large set of unlabeled documents using the Generative Pseudo Labeling approach. They also experimented with generating new search queries.

LIA submitted a total of 5 runs as baselines for Task 1. All five have been included in the pool of results for q-rel evaluation. The first three runs (elastic, meili, and boolean) used bag-of-words models and sparse vector document representation. One was generated with the Elastic 7 Search API (elastic), another was gathered with the Meilisearch system[5] (meili) based on bucket sort, and the third one was produced with a simple boolean model (boolean) powered by PostgreSQL GIN text indexing. Two additional runs relied on embedding vectors based on the paragraph cross-encoder MS MARCO Mini LM[6]. These embeddings, along with a search API based on them, have been released to participants. Documents are ranked based on the dot product between the query and the abstract (vir_abstract) or the title (vir_title) using the pg_vector[7] PostgreSQL extension and an ivvflat dense vector index (k-means vector clustering with $\sqrt{|D|}$ centroids).

Ruby. This team (No paper received) submitted a total of 1 run for Task 1. Their approach relies on ElasticSearch and a TF-IDF score.

Arampatzis. This team (No paper received) submitted a total of 9 runs for Task 1. As these reports are very close, the Tables below only report their evaluation made on their first run.

2.3 Test Results

Table 2 still shows relevance evaluation, with a ranking by NDCG@10 on queries absent from the training qrels. Note that for the sake of brevity, we only consider here the relevance score provided by participants. The selection of top results against the combined score would have modified the relevance evaluation.

2.4 Analysis

We complement these evaluations by taking into consideration other aspects essential for Task 1. Table 3 highlights credibility and text complexity. We used simple automatic metrics to provide an overview of the importance and the complexity of the article. First,

[5] https://www.meilisearch.com/.
[6] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-12-v2.
[7] https://github.com/pgvector/pgvector.

Table 2. Results for CLEF 2024 SimpleText Task 1 on the Test qrels (G01.C1-G10.C1 and T06-T11).

Run	MRR	Precision		NDCG		Bpref	MAP
		10	20	10	20		
AIIRLab_Task1_LLaMABiEncoder	0.9444	0.8167	0.5517	0.6170	0.5166	0.3559	0.2304
AIIRLab_Task1_LLaMAReranker2	0.9300	0.7933	0.5417	0.5943	0.5004	0.3495	0.2177
AIIRLab_Task1_LLaMAReranker	0.8944	0.7967	0.5583	0.5889	0.5011	0.3541	0.2200
LIA_vir_title	0.8454	0.6933	0.4383	0.5013	0.3962	0.3594	0.1534
AIIRLab_Task1_LLaMACrossEncoder	0.7975	0.6933	0.5100	0.4745	0.4240	0.3404	0.1970
LIA_vir_abstract	0.7683	0.6000	0.4067	0.4207	0.3504	0.3857	0.1603
UAms_Task1_Anserini_rm3	0.7878	0.5700	0.4350	0.3924	0.3495	0.4010	0.1824
UAms_Task1_Anserini_bm25	0.7187	0.5500	0.4883	0.3750	0.3707	0.3994	0.1972
UAms_Task1_CE1K_CAR	0.5950	0.5333	0.4583	0.3672	0.3618	0.2701	0.1605
UAms_Task1_CE1K	0.5950	0.5333	0.4583	0.3672	0.3618	0.4032	0.1939
UAms_Task1_CE100_CAR	0.6618	0.5300	0.4567	0.3654	0.3549	0.2657	0.1579
UAms_Task1_CE100	0.6618	0.5300	0.4567	0.3654	0.3549	0.2657	0.1579
AIIRLAB_Task1_CERRF	0.7264	0.5033	0.4000	0.3584	0.3239	0.2204	0.1309
Arampatzis_1.GPT2_search	0.6986	0.5100	0.2550	0.3516	0.2462	0.0742	0.0577
UBO_Task1_TFIDFT5	0.7132	0.4833	0.3817	0.3474	0.3197	0.2354	0.1274
LIA_bool	0.7242	0.5233	0.3633	0.3381	0.2891	0.2661	0.1199
Elsevier@SimpleText_task_1_run8	0.7123	0.4533	0.3367	0.3146	0.2752	0.1582	0.0906
Elsevier@SimpleText_task_1_run4	0.6162	0.4300	0.3217	0.3063	0.2681	0.1642	0.1005
Elsevier@SimpleText_task_1_run10	0.5117	0.4067	0.2767	0.2885	0.2365	0.1236	0.0729
LIA_elastic	0.6173	0.3733	0.2900	0.2818	0.2442	0.3016	0.1325
AB/DPV_SimpleText_task1_FKGL	0.6173	0.3733	0.2900	0.2818	0.2442	0.1966	0.1078
Ruby_Task_1	0.5470	0.4233	0.3533	0.2756	0.2671	0.1980	0.1110
LIA_meili	0.6386	0.4700	0.2867	0.2736	0.2242	0.2377	0.0833
Elsevier@SimpleText_task_1_run6	0.5333	0.3833	0.3117	0.2633	0.2430	0.1841	0.0973
Tomislav/Rowan/Rowan_SimpleText_T1_1	0.5444	0.3733	0.2750	0.2443	0.2183	0.0963	0.0601
Elsevier@SimpleText_task_1_run5	0.4867	0.3533	0.2883	0.2408	0.2232	0.1834	0.0943
Elsevier@SimpleText_task_1_run1	0.5589	0.3000	0.3300	0.2247	0.2399	0.1978	0.1018
Elsevier@SimpleText_task_1_run7	0.4026	0.3200	0.2250	0.2168	0.1850	0.1085	0.0565
Elsevier@SimpleText_task_1_run9	0.3868	0.3300	0.2283	0.2105	0.1829	0.1103	0.0590
Elsevier@SimpleText_task_1_run3	0.4733	0.2367	0.2033	0.1853	0.1703	0.1587	0.0714
Elsevier@SimpleText_task_1_run2	0.4193	0.2233	0.2433	0.1803	0.1865	0.1768	0.0820
Sharingans_Task1_marco-GPT3	0.6667	0.0667	0.0333	0.1149	0.0797	0.0107	0.0107
Tomislav/Rowan_SimpleText_T1_2	0.0217	0.0233	0.0150	0.0121	0.0106	0.0062	0.0025
Petra/Regina_simpleText_task_1	0.0026	0.0000	0.0050	0.0000	0.0035	0.0031	0.0007

the average number of bibliographic references among the top 10 results of each query is provided. Second, we give the average size of sentences and the average number of syllabi per word in the abstract of these results. Note that this time we considered the

Table 3. Evaluation of complexity and credibility for SimpleText Task 1 (over all 176 queries).

Run	Avg #Refs	Avg sentence length	Avg syllabus per word
AB/DPV_SimpleText_task1s_FKGL	9.7	30.4	1.987
Arampatzis_1.GPT2_searchs	10.5	22.1	1.916
LIA_bool	13.0	33.1	1.906
LIA_elastic	9.2	21.1	1.812
LIA_meili	9.6	23.8	1.740
LIA_vir_abstract	7.2	21.0	1.885
LIA_vir_title	9.8	22.4	1.870
AIIRLAB_Task1_CERRF	10.6	22.0	1.895
AIIRLab_Task1_LLaMABiEncoder	9.5	31.0	1.865
AIIRLab_Task1_LLaMACrossEncoder	10.0	30.6	1.890
AIIRLab_Task1_LLaMAReranker	8.8	22.1	1.772
AIIRLab_Task1_LLaMAReranker2	8.6	20.9	1.707
Elsevier@SimpleText_task_1_run1	10.0	22.2	1.888
Elsevier@SimpleText_task_1_run10	10.2	22.3	1.881
Elsevier@SimpleText_task_1_run2	11.2	22.4	1.893
Elsevier@SimpleText_task_1_run3	9.7	22.0	1.894
Elsevier@SimpleText_task_1_run4	10.7	22.2	1.881
Elsevier@SimpleText_task_1_run5	11.1	22.3	1.886
Elsevier@SimpleText_task_1_run6	11.2	22.5	1.885
Elsevier@SimpleText_task_1_run7	9.8	22.2	1.870
Elsevier@SimpleText_task_1_run8	10.3	22.4	1.903
Elsevier@SimpleText_task_1_run9	10.6	22.2	1.857
Petra/Reginas_simpleText_task_1	5.5	22.1	1.955
Ruby_Task_1	9.6	21.2	1.837
Sharingans_Task1_marco-GPT3	9.8	23.0	1.896
Tomislav/Rowan_SimpleText_T1_1	11.3	24.6	1.952
Tomislav/Rowan_SimpleText_T1_2	11.8	23.6	1.943
UAms_Task1_Anserini_bm25	11.8	23.7	1.893
UAms_Task1_Anserini_rm3	11.9	24.8	1.894
UAms_Task1_CE100	11.1	22.3	1.901
UAms_Task1_CE100_CAR	10.6	19.6	1.832
UAms_Task1_CE1K	10.8	22.4	1.872
UAms_Task1_CE1K_CAR	10.2	19.5	1.809
UBO_Task1_TFIDFT5	10.3	22.2	1.899

comb score to elect the top 10 results, which favors the systems w.r.t. the relevance score which may be an orthogonal criterion.

This concludes the results for the CLEF 2024 SimpleText Task 1: Content Selection on Retrieving Passages to Include in a Simplified Summary. Our main findings are the following: First, the Tables on relevance are dominated by neural rankers, in particular, cross-encoders and LLaMA 3 used as a pairwise re-ranker. Second, a majority of participants relied on ElasticSearch search results. If neural models used in processing steps leveraged these results, other IR systems turned out to be competitive. For instance, LIA_vir_title operating with embedding sentences or UAms_Task1_Anserini_rm3, using an Anserini index have high relevance evaluations. Third, as expected, ranking over systems differs according to the considered criterion. Runs filtered against readability measures tend to have shorter sentences with a more or less drop in relevance. Remarkably, LLaMA 3 used as a reranker seems to not only help to select more relevant documents but also with more concise sentences. We refer the reader to the CLEF 2024 SimpleText Task 1 Overview paper [35] for further details and discussion.

3 Task 2: Identifying and Explaining Difficult Concepts

This section details *Task 2: Complexity Spotting* on Identifying and Explaining Difficult Concepts.

3.1 Description

The goal of this task is to decide which concepts in scientific abstracts require explanation and contextualization in order to help a reader understand the scientific text. Since 2023, we have asked participants to identify such concepts and to provide useful and understandable explanations for them. Thus, the task has three subtasks: Task 2.1) to identify candidate terms in a given passage from a scientific abstract, and set the level of difficulty of the concept designated by each term (easy, medium, or difficult); Task 2.2) to provide a definition or an explanation or both only for the difficult terms[8]; Task 2.3) to rank a set of provided definitions for each difficult term.

Data. The corpus of Task 2 is based on the sentences in high-ranked abstracts to the requests of Task 1 collected in 2023 [11]. A total of 175 documents and 1,077 sentences were used to generate the training and test data. In particular, we had 115 documents and 576 sentences for building the training set and 60 documents and 501 sentences for building the test set.

For the training set, we asked 21 experts to manually annotate each training document in order to produce the set of terms for each sentence, the corresponding difficulty, and the definitions and explanations for each difficult term. A total of 1,609 terms and 899 definitions and explanations were generated. In some cases, we intentionally assigned the same documents to different experts in order to have the possibility to further study the agreement between the extraction of terms and the generation of definitions. For each term with a definition, we also generated two "good" definitions and two "bad" definitions in order to create the set of definitions that should be ranked in Task 2.3. A total of 2,356 sentences (equally distributed between good and bad) were generated. In addition to this first set of training data, we also created an additional set of files that were produced by an external expert who reviewed the work of the 21

[8] Henceforth, we will use 'difficult term' to indicate a term that designates a difficult concept.

experts. We called these additional files the validation set. In these files, the external experts added what she thought was missing in the first round of annotation (either terms or definitions or both). An additional set of 677 terms and, 960 definitions, and 3,732 generated definitions (equally distributed between good and bad definitions) were added to the training set.

For the test set, we asked the external expert to annotate the remaining 60 documents. A total of 1440 terms were extracted and 424 definitions were written from the 501 sentences of the test set. An additional 3,816 definitions (equally distributed between good and bad definitions) were also added.

Finally, we encouraged participants to train on existing datasets extracted from other resources such as the WCL dataset [29] to train the definition generation model or use gazetteers, wikification resources as well as resources for abbreviation deciphering.

Evaluation. We evaluated difficult concept spotting, Task 2.1, and their definitions, Task 2.2, in terms of recall, precision, and BLEU score [11]. The ranking of definitions, Task 2.3, will be evaluated with precision@1 and precision@5. A qualitative analysis will also be performed in order to study the problems of term identification and the generation of definitions.

In addition, we will manually evaluate the provided explanations in terms of their usefulness with regard to a query as well as their complexity for a general audience. Note that the provided explanations can have different forms, e.g. abbreviation deciphering, examples, use cases, etc.

3.2 Participant's Approaches

A total of 13 teams submitted 46 runs in total, with many being a combined Task 2.1 and Task 2.2 submission.

AB/DPV. Varadi and Bartulović [40] submitted a total of one run for Task 2. Their main approach was to use natural language processing to extract difficult terms from passages, followed by generating definitions for them or retrieving them from sources such as Wikipedia. The participants did not submit runs on the test set.

AIIRLab. Largey et al. [25] submitted a total of three runs for Task 2. The participants used LLaMA 3 and Mistral language models to create the three runs. The methodology involves prompt engineering and reinforcement learning with human feedback to improve the quality of outputs generated by the LLaMA model.

Dajana/Kathy submitted a total of one run for Task 2. The participants used the LLaMA-2 13B model but they did not provide additional information about the run.

Frane/Andrea submitted a total of one run for Task 2. Participants did not provide additional information about the run.

Sharingans. Ali et al. [2] submitted a total of one run for Task 2. The participants fine-tuned GPT 3.5 turbo model for the selection of difficult terms as well as the generation of definitions and explanations for the extracted terms. Prompt-engineering techniques were employed to construct specific prompts that guided the model in producing accurate and contextually relevant definitions.

SINAI. Ortiz-Zambrano et al. [33] submitted a total of three runs for Task 2. Their approach is to apply learning cues without prior examples to the GPT-4-Turbo model, extracting predictions from the generated sequence. They also used the OpenAI API in Python to interact with the model, allowing for an easy integrate GPT-4-Turbo into their workflow.

Petra/Regina. Elagina and Vučić [10] submitted a total of one run for Task 2. The participants employed a combination of named entity recognition (NER) techniques and rule-based approaches to identify and extract entities such as proteins, genes, and chemical compounds. In particular, they used spaCy for NER and developed custom rules for entity extraction.

Tomislav/Rowan. Mann and Mikulandric [27] submitted a total of two runs for Task 2. The team created a prompt for LLaMA 2 13B model that asked the LLM to iterate over each of their source sentences and extract three scientific terms from each sentence and then were sorted and prompted to LLaMA to return a difficulty rating. Wikipedia was used to return definitions for the difficult terms.

UAmsterdam. Bakker et al. [4] submitted a total of three runs for Task 2. The participants used an IDF-based term weighting to locate the rarest terms for Task 2.1. For Task 2.3, they developed an approach to rank definitions or explanations for a given sentence and term pair looking at the textual similarity of the large set of provided sentences.

UBO. Vendeville et al. [41] submitted a total of one run for Task 2. In particular, the participants used a Small Language Model, Phi3 mini, without fine-tuning with a one-shot prompt.

UniPD. Di Nunzio et al. [8] submitted a total of three runs for Task 2. Their participation in Task 2 focused on identifying and explaining difficult content using Large Language Models (LLMs) to enhance text simplification. The methodology involves iterative experimentation with various prompting strategies to optimize the performance of the model in this task.

Ruby. (No paper received) submitted a single run for Task 2.1. This run can only be evaluated on the train data.

Arampatzis. (No paper received) made 5 submissions for Tasks 2.1 and 2.2 each, and two submissions for Task 2.3. Their Task 2.1 and Task 2.2 submissions only contain results for the train data.

3.3 Results

In this section, we present the results on the test set for Task 2.1 and Task 2.2. At present time, the results for Task 2.3 are still ongoing (with only four runs by two participants) and will be made available in the future.

3.4 Test Results

The results on the test set are summarized in Table 4. For each run, we report:

- the recall of all the terms, independently from the level of difficulty;
- the precision of all the terms, independently from the level of difficulty;
- the recall of the difficult terms;
- the precision of the difficult terms;
- the BLEU score computed for bigrams (ngrams with n = 2).[9]

Table 4. Results for CLEF 2024 SimpleText Task 2

runid	recall	precision	recall_difficult	The results shown in	bleu_n2_average
AIIRLab_Task2.2_LLaMA	0.28	0.65	0.26	0.67	0.15
AIIRLab_Task2.2_LLaMAFT	0.01	0.99	0.00	1.00	0.12
AIIRLab_Task2.2_Mistral	0.41	0.69	0.19	0.49	0.13
Dajana/Kathy_SimpleText_Task2.2_LLaMA2_13B_CHAT	0.01	0.59	0.00	0.00	0.00
Frane/Andrea_SimpleText_Task2.2_LLaMA2_13B_CHAT	0.01	0.65	0.01	0.36	0.00
team1_Petra_and_Regina_Task2_ST	0.00	0.50	0.00	0.00	0.00
Sharingans_Task2.2_GPT	0.47	0.55	0.54	0.60	0.10
SINAI_task_2_PRM_ZS_TASK2_V1	0.09	0.82	0.10	0.52	0.16
SINAI_task_2_PRM_ZS_TASK2_V2	0.16	0.78	0.13	0.77	0.16
SINAI_task_2_PRM_ZS_TASK2_V3	0.10	0.86	0.05	0.83	0.11
Tomislav/Rowan_Task2.2_LLaMA2_13B_CHAT_1	0.01	0.61	0.00	0.00	0.00
Tomislav/Rowan_Task2.2_LLaMA2_13B_CHAT	0.01	0.33	0.00	0.00	0.00
UAms_Task2-1_RareIDF	0.09	0.97	0.03	0.09	0.00
UboNLP_Task2.1_phi3-oneshot	0.54	0.65	0.32	0.37	0.00
unipd_t21t22_chatgpt	0.13	0.64	0.08	0.62	0.19
unipd_t21t22_chatgpt_mod1	0.22	0.52	0.20	0.60	0.18
unipd_t21t22_chatgpt_mod2	0.31	0.46	0.34	0.69	0.01

We also show the trend of the main results in Fig. 1. In the top row, the recall and precision graphs for all the terms and only the difficult terms are shown. In the bottom row, the BLEU score of the proposed definitions in relation with the recall and precision on difficult terms.

3.5 Analysis

The results shown in the previous subsection reveal that the use of large language models for the extraction of terms, the assessment of the difficulty of these terms, and the generation of the definitions to explain the difficult concepts are at an initial stage that will open new perspective in the Automatic Term Extraction panorama. In particular, compared to the recent results and surveys [see 7], the values of recall and precision are sufficiently good but suboptimal when compared to the state-of-the-art models (of course, we need to take into consideration that this is the first time participants dealt with this new dataset).

This concludes the results for the CLEF 2024 SimpleText Task 2: Complexity Spotting on Identifying and Explaining Difficult Concepts. Our main findings are the following: First, the runs submitted by the participants to this task are quite stable in terms

[9] https://cran.r-project.org/web/packages/sacRebleu/vignettes/sacReBLEU.html.

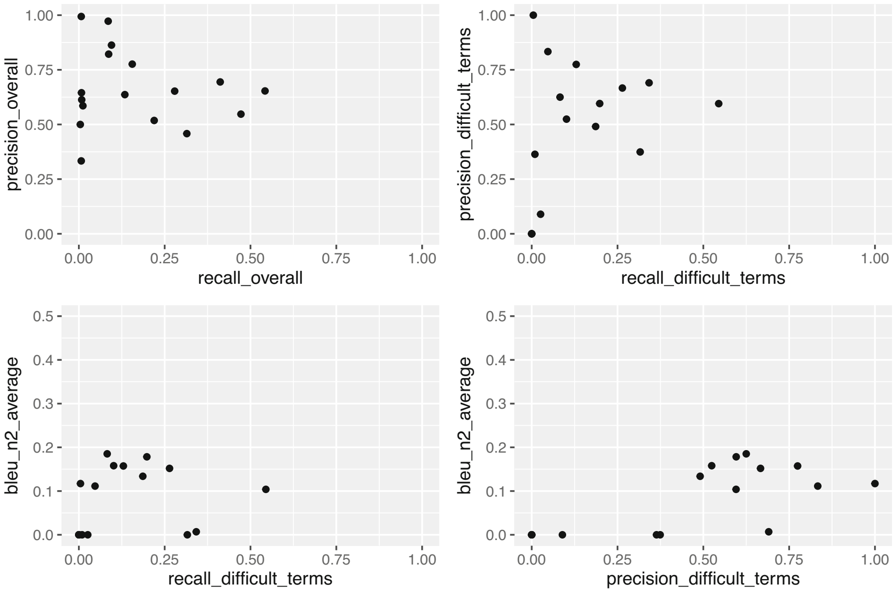


Fig. 1. Summary of Task 2 results.

of recall-precision performances when dealing with all the terms or the difficult ones. Independently from the difficulty of terms, the models proposed by the participants can achieve precision higher than .50 across a range of recall values. Second, the BLEU score of the generated definitions is also relatively stable ranging from 0.1 to 0.2 for any recall and precision values. Third, the best-performing runs are usually those that have some analysis of the optimal prompting or a manual interaction with the model. This is in line with the latest research studies on this issue [26]. We refer the reader to the CLEF 2024 SimpleText Task 2 Overview paper [31] for further details and discussion.

4 Task 3: Simplify Scientific Text

This section details *Task 3: Text Simplification* on Simplify Scientific Text.

4.1 Description

The goal of this task is to provide a simplified version of the sentences extracted from scientific abstracts. Participants will be provided with popular science articles and queries and matching abstracts of scientific papers, either split into individual sentences or as the entire abstracts.

Data. Task 3 uses the same corpus based on the sentences in high-ranked abstracts to the requests of Task 1. Our training data is a truly parallel corpus of directly simplified sentences (893 sentences this year) coming from scientific abstracts from the DBLP

Citation Network Dataset for *Computer Science* and Google Scholar and PubMed articles on *Health and Medicine*.

Other existing text simplification corpora used post-hoc aligned sentences [42, e.g.,]. The SimpleText corpus contains 893 directly simplified sentences, and a useful addition to existing high-quality corpora like NEWSELA [42] (2,259 sentences). Our track is the first to focus on the simplification of scientific text with a much higher text complexity than news articles.

Table 5. Example of SimpleText Task 3 reference versus input: ~~deletions~~ and insertions

Topic	**Document**	**Output**
G01.1	130055196	~~As various kinds~~ The rise of output devices ~~emerged , such as highresolution~~ like high-resolution printers ~~or a display of~~ and PDA ~~(Personal Digital Assistant) ,~~ displays has increased the ~~importance of~~ need for high-quality resolution conversion ~~has been increasing~~ . ~~This~~ The paper proposes a new method ~~for enlarging image with~~ to make images bigger while maintaining high quality . ~~One of the largest problems on image enlargement~~ The main issue with enlarging images is ~~the exaggeration of the jaggy~~ that jagged edges can become exaggerated . To ~~remedy~~ solve this problem , we ~~propose~~ suggest a new ~~interpolation~~ method ~~, which uses artificial~~ that helps us to estimate the value of the newly generated pixels using a neural network ~~to determine the optimal values of interpolated pixels~~ . The ~~experimental~~ experiment 's results are ~~shown~~ presented and ~~evaluated~~ analyzed . ~~The~~ We evaluate the effectiveness of our methods ~~is discussed~~ by comparing ~~with the conventional methods~~ them to traditional approaches .

Table 5 shows an example of a human reference simplification, combining the input sentences belonging to the abstract of the document $id = 130055196$ retrieved for query G01.1. Here, we show the deletions and insertions relative to the source input sentences (in this case in the first 4 sentences).

Available training data from 2023 includes 983 sentences from scientific abstracts plus manual simplifications [13]. These text passages were simplified either by master students in Technical Writing and Translation or by a domain expert (a computer scientist) and a professional translator (native English speaker) working together. The new sentence-level evaluation (test) data in 2024 consists of 578 sentences.

In 2024, we expanded the training and evaluation data. In addition to sentence-level text simplification, we will provide passage-level input and reference simplifications, with the train data corresponding to 175 abstracts with corresponding human simplifications. The new abstract-level evaluation (test) data in 2024 consists of 103 abstracts.

Evaluation. In 2024, we emphasize large-scale automatic evaluation measures (SARI, BLEU, compression, readability) that provide a reusable test collection. This automatic evaluation will be supplemented with a detailed human evaluation of other aspects, essential for deeper analysis. Almost all participants used generative models for text simplification, yet existing evaluation measures are blind to potential hallucinations with extra or distorted content [13]. In 2024, we provide further analysis of ways to detect and quantify spurious content in the output, potentially corresponding to what is informally called "hallucinations".

4.2 Participant's Approaches

A total of 15 teams submitted 83 runs in total.

AB/DPV. Varadi and Bartulović [40] submitted one run for Task 3. Their approach is an LSTM model for the sentence-level task.

Sharigans. Ali et al. [2] submitted a total of two runs for Task 3. Their approach is a GPT-3.5 model for both the sentence-level and abstract-level tasks.

Tomislav/Rowan. Mann and Mikulandric [27] submitted a total of two runs for Task 3. Their approach is the LLama 2 model with a range of prompts and post-processing for both the sentence-level and abstract-level tasks.

Petra/Diana. Elagina and Vučić [10] submitted one run for Task 3. Their approach is a LLaMA model for the sentence-level task.

Dajana/Katya submitted one run for Task 3. Their approach which follows standard text simplification approaches is applied to the sentence-level task.

AIIRLab Largey et al. [25] submitted a total of eight runs for Task 3. Their approach uses LLaMA3 and Mistral models with different prompting and fine-tuning, for both the sentence-level and abstract-level tasks.

UBO Vendeville et al. [41] submitted a total of four runs for Task 3. Their approach is to prompt a smaller Phi3 model for lexical and grammatical text simplifications, for both the sentence-level and abstract-level tasks.

UAmsterdam Bakker et al. [4] submitted a total of ten runs for Task 3. They experiment with GPT-2, and Wiki and Cochrane-trained models at the sentence, paragraph, and document-level text simplification, for both sentence-level and document-level tasks.

UZHPandas Michail et al. [28] submitted a total of ten runs for Task 3. They experiment with a multi-prompt Minimum Bayes Risk (MBR) decoding approach to the sentence-level task.

Elsevier Capari et al. [5] submitted a total of ten runs for Task 3. Their approach is based on a GPT-3.5 model experimenting with zero-shot and few-shot prompts for both sentence-level and abstract-level tasks.

Frane/Andrea submitted one run for Task 3. Their approach which follows standard text simplification approaches is applied to the sentence-level task.

Arampatzis. (No paper received) submitted a total of eight runs for Task 3. Their approach is a range of models (DistilBERT, T5) for both the sentence-level and abstract-level tasks.

Ruby. (No paper received) submitted two runs for Task 3. Their approach uses standard models for both sentence-level and abstract-level tasks.

SONAR. (No paper received) submitted a single run for Task 3. Their approach is a standard model for the sentence-level task.

PiTheory. (No paper with run details received) submitted a total of twenty runs for Task 3. Their approach uses pre-trained BART and T5 models but contains very few results for both the sentence-level and abstract-level tasks.

4.3 Results

This section details the results of the task, for both sentence-level and abstract-level test simplification subtasks.

Table 6. Results for CLEF 2024 SimpleText Task 3.1 sentence-level text simplification (task number removed from the run_id) on the test set

run_id	count	FKGL	SARI	BLEU	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
Source	578	13.65	12.02	19.76	1	1	1	1	0	0	8.8
Reference	578	8.86	100	100	0.7	1.06	0.6	0.01	0.27	0.54	8.51
Elsevier_run1	578	10.33	43.63	10.68	0.87	1.06	0.59	0.00	0.45	0.53	8.39
Elsevier_run4	577	11.73	43.14	12.08	0.85	1.00	0.63	0.00	0.37	0.50	8.54
Elsevier_run8	577	12.40	42.95	12.35	0.90	1.02	0.63	0.00	0.35	0.50	8.66
Elsevier_run6	577	12.65	42.88	11.76	0.95	1.00	0.64	0.00	0.38	0.47	8.63
Elsevier_run7	577	12.55	42.87	12.20	0.87	1.00	0.63	0.00	0.35	0.51	8.67
Elsevier_run9	577	12.53	42.61	12.15	0.87	1.00	0.63	0.00	0.35	0.50	8.67
Elsevier_run3	577	11.50	42.58	15.75	0.76	0.98	0.68	0.00	0.23	0.46	8.68
Elsevier_run10	577	12.57	42.49	11.91	0.91	1.02	0.63	0.00	0.34	0.50	8.67
AIIRLab_llama-3-8b_run1	578	8.39	40.58	7.53	0.90	1.37	0.56	0.00	0.48	0.58	8.45
AIIRLab_llama-3-8b_run3	578	9.47	40.36	6.26	1.17	1.52	0.53	0.00	0.53	0.56	8.51
AIIRLab_llama-3-8b_run2	578	10.33	39.76	5.46	1.03	1.19	0.51	0.00	0.60	0.56	8.34
UZHPandas_simple_cot	578	13.74	39.59	3.38	3.44	2.67	0.41	0.00	0.76	0.12	8.61
UZHPandas_simple	578	11.24	39.28	5.67	0.88	0.98	0.52	0.00	0.53	0.62	8.45
Sharingans_finetuned	578	11.39	38.61	18.18	0.83	1.07	0.77	0.11	0.16	0.32	8.70
UZHPandas_selection_sle_cot	578	6.49	38.38	1.03	4.76	6.26	0.30	0.00	0.89	0.14	8.30
UZHPandas_simple_inter_def	578	21.36	38.29	3.13	1.93	0.99	0.46	0.00	0.69	0.33	8.86
UZHPandas_selection_lens_cot	578	6.74	38.16	1.10	4.54	5.88	0.32	0.00	0.87	0.14	8.32
UZHPandas_5Y_target_cot	578	6.39	37.95	0.97	4.73	6.25	0.30	0.00	0.89	0.14	8.30
UZHPandas_selection_lens	578	21.29	37.79	2.71	1.97	1.01	0.44	0.00	0.71	0.34	8.85
UBO_Phi4mini-s	578	8.74	36.78	0.58	18.23	23.48	0.47	0.00	0.66	0.29	8.89
UZHPandas_selection_lens_1	578	7.79	36.72	3.65	0.72	0.98	0.46	0.00	0.54	0.73	8.25
UBO_Phi4mini-sl	578	6.16	36.53	0.61	6.92	9.81	0.38	0.00	0.80	0.42	8.72
UZHPandas_5Y_target_inter_def	578	19.30	36.53	2.27	1.76	1.01	0.45	0.00	0.70	0.41	8.87
UZHPandas_selection_sle	578	6.07	35.30	2.57	0.65	0.98	0.43	0.00	0.56	0.78	8.17
UZHPandas_5Y_target	578	5.94	34.91	2.29	0.66	0.99	0.43	0.00	0.57	0.78	8.17
RubyAiYoungTeam	578	8.76	34.40	15.37	0.60	1.22	0.69	0.03	0.05	0.44	8.71
SONAR_SONARnonlinreg	578	13.14	32.12	18.41	0.97	1.01	0.93	0.13	0.11	0.13	8.73
UAms_GPT2_Check	578	11.47	29.91	15.10	1.02	1.23	0.87	0.14	0.17	0.14	8.68
UAms_GPT2	578	10.91	29.73	13.07	1.30	1.50	0.79	0.06	0.29	0.12	8.63
Arampatzis_T5	578	13.18	28.92	10.66	1.12	1.10	0.72	0.03	0.34	0.37	9.06
UAms_Wiki_BART_Snt	578	12.13	27.45	21.56	0.85	0.99	0.89	0.32	0.02	0.16	8.73
Arampatzis_DistilBERT	578	5.85	19.00	13.56	1.03	3.00	0.95	0.00	0.22	0.11	8.65
UAms_Cochrane_BART_Snt	578	13.22	18.45	19.21	0.95	0.99	0.96	0.59	0.02	0.07	8.77
Arampatzis_METHOD	578	13.65	12.12	19.77	1.00	1.00	1.00	0.99	0.00	0.00	8.80

4.4 Task 3.1: Sentence-Level Scientific Text Simplification

Table 6 shows the Task 3.1 (sentence-level text simplification) results. The table is restricted to submissions covering a sufficient number of input sentences. We show a number of evaluation scores against the human reference simplifications, in particular SARI and BLEU. In addition, we provide additional text statistics on the system output such as FKGL, and a comparison to the source input.

We make a number of observations. First, the table is sorted on SARI, the main automatic text simplification measure used in the track. We observe SARI scores of 30+ % for the majority of systems and 40+ % for the top-scoring systems. This high overlap with the human reference simplifications is encouraging and indicates that the effectiveness of text simplification approaches, traditionally trained on youth news reading corpora like Newsela, also extends to scientific text.

Second, in terms of the level of text complexity, readability measures like FKGL provide a rough indicator of lexical and grammatical complexity. The original sentences have an FKGL of 13–14 corresponding to university-level text, and the majority of systems reduce this to an FKGL of 11–12 corresponding to the exit level of compulsory education. This is an encouraging result, as it indicates that the scientific text simplification approach can be a viable approach to lower the textual complexity of scientific text toward the range acceptable by a layperson. Although this is positive indicator, this approximate measure does not take into account terminological complexities as studied in Task 2, or ways to retrieve all and only more accessible abstracts in Task 1 [14].

Third, the table includes various other scores that indicate that there is still considerable room for improvement in scientific text simplification. Throughout the table the BLEU evaluation measure remains very low, and leads to a different ranking of systems with some of the best systems on BLEU demonstrating superior overlap with the human reference simplifications. The table also reveals some runs with very high "compression" ratios and sentence splits, as well as high proportions of additions. While evaluation measures like SARI are essential for understanding important aspects of text simplification output quality, they are also known to be relative insensitive to content outside the intersection with the manual text simplifications. Hence high levels of insertion of content can still lead to favorable SARI scores, and even improve text statistics like FKGL, without conveying key content of the original text.

4.5 Task 3.2: Abstract-Level Scientific Text Simplification

Table 7 shows the Task 3.2 (abstract-level text simplification) results. Again we restrict the table to submissions covering a sufficient number of input abstracts.

Table 7. Results for CLEF 2024 SimpleText Task 3.2 abstract-level text simplification (task number removed from the run_id) on the test set

run_id	count	FKGL	SARI	BLEU	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
Source	103	13.64	12.81	21.36	1	1	1	1	0	0	8.88
Reference	103	8.91	100	100	0.67	1.04	0.6	0	0.23	0.53	8.66
AIIRLab_llama-3-8b_run1	103	9.07	43.44	11.73	1.01	1.38	0.51	0.00	0.37	0.56	8.57
AIIRLab_llama-3-8b_run3	103	10.17	43.21	11.03	1.15	1.47	0.52	0.00	0.40	0.51	8.66
Elsevier_run2	103	11.01	42.47	10.54	1.04	1.22	0.51	0.00	0.38	0.55	8.60
AIIRLab_llama-3-8b_run2	103	10.22	42.19	7.99	1.31	1.38	0.48	0.00	0.53	0.52	8.44
Elsevier_run5	103	12.08	42.15	10.96	1.04	1.15	0.52	0.00	0.36	0.53	8.75
Sharingans_finetuned	103	11.53	40.96	18.29	1.20	1.39	0.65	0.00	0.24	0.34	8.80
UBO_Phi4mini-ls	103	8.45	38.79	5.53	1.21	1.75	0.43	0.00	0.40	0.63	8.53
UBO_Phi4mini-l	103	9.96	38.41	10.01	1.29	2.11	0.55	0.00	0.24	0.51	9.03
UAms_GPT2_Check_Abs	103	12.85	36.47	13.12	0.91	0.92	0.59	0.00	0.18	0.45	8.73
UAms_Cochrane_BART_Doc	103	14.46	33.51	9.39	0.65	0.58	0.54	0.04	0.06	0.53	8.80
UAms_Cochrane_BART_Par	103	16.53	31.58	15.40	1.08	0.80	0.67	0.04	0.15	0.32	8.81
UAms_GPT2_Check_Snt	103	11.57	30.71	15.24	1.54	1.70	0.78	0.00	0.27	0.13	8.77
Arampatzis_METHOD	103	0.00	28.28	0.00	0.00	0.00	0.00	0.00	0.00	1.00	10.82
Arampatzis_T5	103	0.00	28.28	0.00	0.00	0.00	0.00	0.00	0.00	1.00	10.82
Arampatzis_DistilBERT	103	0.00	28.28	0.00	0.00	0.00	0.00	0.00	0.00	1.00	10.82
UAms_Wiki_BART_Doc	103	15.68	26.50	15.11	1.51	1.14	0.76	0.01	0.25	0.11	8.79
UAms_Wiki_BART_Par	103	13.11	23.92	19.49	1.39	1.37	0.81	0.01	0.11	0.10	8.86

We make a number of observations. First, in terms of evaluation measures like SARI we see again similar encouraging performance levels when evaluating against the human reference simplifications. This is partly due to the use of proven sentence-level text simplification models with the output merged back into the entire abstract. Second, there remains room for improvement in capturing the human simplifications more closely, as the BLEU score remains low throughout. Here, the more conservative approaches seem to obtain better scores. Third, we see less extreme values on the other indicators, but still considerable variation in the compression ratio and number of splits, and proportions of addition and deletions. We will investigate how much of the output is grounded in the source sentences and abstracts below.

Many submissions rely on proven sentence-level text simplification approaches, with results closely mirroring those observed for the sentence-level task. It is encouraging to see solid performance for the approaches that perform text simplification at the entire abstracts in one pass. This holds the promise to incorporate the discourse structure, use more complex text simplifications operations such as deletions and merges, and deploy planner-based approaches to the text simplification of long documents.

4.6 Analysis

We conduct a deeper analysis of how much of the generated simplified output sentences and abstracts can be traced to the source input. In particular, we look at spurious generated content and it's prevalence in the submitted generated text simplifications. This content is at risk of being introduced gratuitously by the generative model, and what is informally referred to as "hallucinations."

Table 8. Example of SimpleText Task 3 output versus input: ~~deletions~~, insertions, and whole sentence insertions

Topic	Document	Output
G01.1	130055196	As various kinds of output devices emerged , such as highresolution printers or a display of PDA (Personal Digital Assistant) ~~, the~~ . The importance of high-quality resolution conversion has been increasing . This paper proposes a new method for enlarging an image with high quality . It will involve using a combination of high-speed imaging and high-resolution video . One of the ~~largest~~ biggest problems on image enlargement is the exaggeration of the jaggy edges . This is especially true when the image is enlarged , as in this case . To remedy this problem , we propose a new interpolation method ~~, which~~ . This method uses artificial neural network to determine the optimal values of interpolated pixels . The experimental results are shown and evaluated . The results are compared to other studies and found to be inconclusive . The effectiveness of our methods is discussed by comparing with the conventional methods . Our methods are designed to help people with mental health problems , not just as a way to cure them .

Earlier in Table 5, we showed an example of a human reference simplification, combining the input sentences belonging to the abstract of the document $id = 130055196$ retrieved for query G01.1. We can do the same for the automatically generated scientific text simplifications. We show again the deletions and insertions relative to the source input sentences. Table 8 shows an example output simplification of one of the participating teams, for the same input sentences as in Table 5 above. Most simplifications are revisions of the input, but we also observe that sometimes an entire sentence is inserted (shown as xxx in Table 8).

We provide a detailed analysis quantifying the prevalence of spurious content in the CLEF 2024 SimpleText Task 3 Overview paper [17]. We re-aligned the generated output with the original source sentences, and flag here only entire output sentences that do not share a single token with the input. The example in Table 8 is an extreme case picked to illustrate both the importance and complexity of detecting such spurious content. However, our analysis reveals that the amount of spurious content is varying but far from infrequent. A total of 17 out 36 submissions (47%) have spurious whole sentences in at least 10% of the input sentences. In fact, 14 (39%) submissions in at least 20% of the input, and 7 (19%) submissions in at least 50% of the input sentences. The detection of non-aligned output sentences is indicative but imperfect. For example, significant reordering of content may lead to false positives in rare cases, and unusual tokenization or formatting may affect the alignment with the source even systematically. Note also that the detected additions may introduce helpful background knowledge or other useful information to contextualize the information in the source sentences.

We make a number of observations based on our analysis in this section. First, the fraction of sentences with spurious content is very low for some submissions, however, for other submissions, the fraction is very substantial. Second, the standard evaluation measures used for text simplification, and in fact for any text generation task in NLP, do not take this aspect into account. A submission with significant spurious content can still obtain very high text overlap with the reference, and hence obtain a very high performance score. Third, and more generally, human evaluation and this type of analysis feel crucial to accurately evaluate generative models for the NLP and IR challenges addressed in our Track and in CLEF in general.

This concludes the results for the CLEF 2024 SimpleText Task 3: Text Simplification on Simplify Scientific Text. Our main findings are the following: First, we observe competitive performance for scientific text simplification, both on evaluation against the human reference simplifications and on text statistics such as FKGL readability score. Second, the abstract-level text simplification results is a mixture of sentence-level and passage-level text simplification approaches. Third, our analysis reveals a very high and varying range of spurious text generation, not detected by standard evaluation measures, and a major concern in the use of these model in a real-world setting. More generally, almost all participants use generative models (for the task, the track, and CLEF in general), and the track offers a unique setting to study some of the inherent limitations of generative models. We refer the reader to the CLEF 2024 SimpleText Task 3 Overview paper [17] for further details and discussion.

5 Task 4: Tracking the State-of-the-Art in Scholarly Publications

This section details *Task 4: SOTA?* on Tracking the State-of-the-Art in Scholarly Publications.

5.1 Description

In Artificial Intelligence (AI), a common research objective is the development of new models that can report state-of-the-art (SOTA) performance. The reporting usually comprises four integral elements: Task, Dataset, Metric, and Score. These (Task, Dataset, Metric, Score) tuples coming from various AI research papers go on to power leaderboards in the community. Leaderboards, akin to scoreboards, traditionally curated by the community, are platforms displaying various AI model scores for specific tasks, datasets, and metrics. Examples of such platforms include the benchmarks feature on the Open Research Knowledge Graph and Papers with Code (PwC). Utilizing text mining techniques allows for a transition from the conventional community-based leaderboard curation to an automated text mining approach. Consequently, the goal of Task 4: SOTA? is to develop systems which given the full text of an AI paper, are capable of recognizing whether an incoming AI paper indeed reports model scores on benchmark datasets, and if so, to extract all pertinent (Task, Dataset, Metric, Score) tuples presented within the paper.

Data. The training and test datasets for this task are derived from community-curated (T, D, M, S) annotations for thousands of AI articles available on PwC (CC BY-SA). We will utilize the dataset obtained from our prior work, specifically the PwC source downloaded on May 10, 2021 [20,23], which comprised over 7,500 articles. These articles, originally sourced from arXiv under CC-BY licenses, are available in TEI XML format, each accompanied by one or more (T, D, M, S) annotations from PwC. While our previous work employed dataset splits for two-fold cross-validation experiments, for the SimpleText Task 4, we will establish new 70/30 train/test splits, providing approximately 5,000 annotated articles for participant training. A preliminary version of our training dataset can be accessed on Github https://github.com/jd-coderepos/sota.

The test set will strategically include only those articles with TDMs seen in the training set, creating a few-shot evaluation setting. Furthermore, in our subsequent research [21], we explored a zero-shot evaluation setting, wherein the dataset contained articles with at least one T, D, or M not seen in the model's training set. Thus in addition to the few-shot evaluation, we intend to introduce a second evaluation setting for Task 4, evaluating models in a zero-shot context, for which a new test dataset will be created. Finally, ongoing efforts involve expanding the primary task corpus by incorporating approximately 1,500 articles into both the train and test sets that do not report leaderboards. These articles will be annotated with the *unknown* label. Consequently, systems developed in our shared task will have comprehensive applicability to any AI article, extracting (T, D, M, S) annotations for articles that contain them and assigning *unknown* for those that do not.

Evaluation. As discussed above, in Task 4 participant systems will be evaluated in the two evaluation settings. For **Few-shot** evaluation, trained systems will have to predict

(T, D, M, S) annotations on a new collection of articles' full-text. The labels in the gold dataset will include only (T, D, M, S)'s seen at least once in training. For **Zero-shot** evaluation, the task is as above with a different collection of articles, which have (T, D, M, S) with unseen T, D, or M in the training set. In both settings, the standard recall, precision, and F-score metrics will be used to report scores to the participant systems.

5.2 Participant's Approaches

A total of 2 teams submitted 36 runs in total.

AMATU by Staudinger et al. [36] submitted a total of three runs for the **few-shot** evaluation phase of Task 4. They submitted nine runs for the **zero-shot** evaluation phase of Task 4. Their general approach to extract the (T, D, M, S) annotations were in two main categories: 1) a pure pattern-based approach inspired after AxCell [24], and 2) an AI-based approach using LLMs with a zero-shot prompt and a few-shot prompt tested for GPT-3.5 [32] and Mistral-7B [19]. For the latter category, they also experimented with variants on the input scholarly article text from which the (T, D, M, S) annotations were expected to be extracted. This we generally refer to as the *context*. They tried two context variants: 1) full paper text and 2) only the text from sections referring to experiments and results, in addition to the abstract, which was pre-extracted inspired by the Argumentative Zoning (AZ) method [38].

L3S by Kabongo et al. [22] submitted a total of 12 runs for the **few-shot** evaluation phase of Task 4. They submitted 12 runs for the **zero-shot** evaluation phase of Task 4. Their approach entailed leveraging the FLAN-T5 [6] strategy which encompassed fine-tuning a pre-trained LLM with a standard set of instructions to better equip them to handle various tasks. Leveraging the applicable instructions from the FLAN-T5 collection, they fine-tuned LLMs, viz. Mistral-7B and LLaMA 2 [39], to make them better suited to handle the (T, D, M, S) extraction task. Furthermore, they also tested the most recent proprietary GPT models viz. GPT-4 [1] and GPT-4o. Finally, as the information extraction context they tried 3 different methods: DocTAET ((T)-title, (A)- abstract, (E)-experimental setup, and (T)-tabular information parts of the full-text), DocREC (text selected from the sections named (R)-results, (E)-experiments, and (C)-conclusions), and DocFULL (full paper text). Resultingly, for each evaluation phase they submitted a total of 4 models x 3 contexts = 12 runs.

5.3 Results

Table 9 and Table 10 present a summary of the results from the two teams. Overall, given *Team AMATU*'s approaches, the pattern-based method proved a competitive solution to the SOTA challenge, in comparison to advanced LLM-based solutions. While the LLM solution did outperform the pattern-based approach the difference was minor. Furthermore, comparing the zero-shot versus few-shot paradigms, the LLMs were significantly more effective in the few-shot setting i.e. when shown successful task completion outputs. Also, the LLM performed significantly better when given the full paper text as input from which to extract (T, D, M, S) as opposed to given selective text using the AZ method.

Table 9. Evaluation results for the binary classification or filtering of papers with and without leaderboards (reported as General Accuracy) and as a structured summary generation task (reported with ROUGE metrics). *Team AMATU*'s few-shot evaluation results are reported for AxCell and their zero-shot evaluation results are reported for GPT-3.5 via the few-shot prompting paradigm. *Team L3S*'s results are reported for Mistral-7B finetuned with the DocTAET context. The best results are shown in bold.

	Few-shot					Zero-shot				
	Rouge				Gen.	Rouge				Gen.
	1	2	L	Lsum	Acc.	1	2	L	Lsum	Acc.
AMATU	**58.34**	12.98	**57.34**	54.4	75.59	**73.72**	6.07	72.72	72.57	85.93
L3S	57.24	**19.67**	56.28	**56.19**	**89.68**	73.54	**12.23**	**73.01**	**72.95**	**95.97**

Table 10. Evaluation results w.r.t. the individual (Task, Dataset, Metric, Score) elements and Overall in terms of **F1 score**. *Team AMATU*'s few-shot evaluation results are reported for AxCell and their zero-shot evaluation results are reported for GPT-3.5 via the few-shot prompting paradigm. *Team L3S*'s results are reported for Mistral-7B finetuned with the DocTAET context. The best results are shown in bold.

Model	Mode	Few-shot					Zero-shot				
		T	D	M	S	Overall	T	D	M	S	Overall
AMATU	Exact	27.11	**23.22**	**24.85**	**9.34**	**21.13**	10.01	13.16	11.65	**9.85**	11.16
	Partial	28.08	24.92	25.8	10.86	22.62	16.12	17.12	13.72	11.1	14.52
L3S	Exact	**33.38**	18.51	24.23	1.87	19.50	**26.99**	**14.32**	**22.04**	1.20	**16.14**
	Partial	46.35	32.75	34.16	2.25	28.88	44.90	27.29	32.23	1.41	26.46

For *Team L3S*, in both the evaluation phases, their model results showed that minimal finetuning of relatively smaller LLMs, specifically Mistral-7B, equips them for (T, D, M, S) extraction task surpassing the performance of LLMs, specifically the latest GPT-4 proprietary models, with a significantly more vast parameter space. The overall best results even for the extraction of the (T, D, M, S) elements was obtained by Mistral given the DocTAET context.

Comparing *Team AMATU* and *Team L3S*, none of the systems from the former team were finetuned to the task. Thus *Team AMATU* presents novel insights into the community to leveraging LLM's effectively for the (T, D, M, S) extraction objective using clever prompt engineering strategies that shows comparable performans to the latter teams' computationally intensive finetuning approach. It maybe that finetuning would be essential to create the most optimal model, however, from the team's solutions the importance prompt engineering for effective downstream performance is clearly emphasized.

This concludes the results for the CLEF 2024 SimpleText Task 4: SOTA? on Tracking the State-of-the-Art in Scholarly Publications. Our main findings are the following: First, effective prompting paradigms should be a go-to strategy to test LLMs out-of-the-box for the SOTA? shared task objective. Second, finetuning small-scale models makes them better able to handle the SOTA? objective than larger-scale LLMs known for their generative AI abilities when simply applied to the IE task. Third, the paper context over which the IE task is expected to be performed must have an ideal balance of length versus selectivity of specific sections in the paper that indeed are highly likely to contain mentions of the (T, D, M, S). On the extreme end of the spectrum, using the full paper text without effective context selection hinders and seems to distract the LLM downstream IE task performance. We refer the reader to the CLEF 2024 SimpleText Task 4 Overview paper [9] for further details and discussion.

6 Conclusions

This paper described the setup of the CLEF 2024 SimpleText track, which contains four interconnected tasks on scientific text summarization and simplification. More detailed discussion on each of the tracks can be found in the task overview papers: Task 1 on *Content Selection* [35]; Task 2 on *Complexity Spotting* [31]; Task 3 on *Text Simplification* [17]; and Task 4 on *SOTA?* [9].

The main aim of our track, and the CLEF evaluation forum as a whole, is i) to construct corpora and evaluation resources to stimulate research on scientific text summarization and simplification, and ii) to foster a community of IR, NLP, and AI researchers working together on the important task of making science more accessible for everyone.

Within the CLEF 2024 SimpleText track, we have constructed extensive corpora and manually labeled evaluation data. First, a large corpus of over 4 million scientific abstracts that can be used for popular science search requests, with corresponding relevance judgments. Second, scientific terms from sentences coming from scientific abstracts with manually attributed difficulty scores and terminology expert provided explanations and definitions. Third, a parallel corpus of manually simplified sentences and abstracts from the scientific literature. Fourth, a corpus for extracting key performance indicators on standard benchmarking data from full text scientific papers. These reusable corpora and evaluation resources are available to participants and other researchers who want to work on the important problem of making scientific information open and easily accessible for everyone.

In terms of a building a community researching scientific text summarization and simplification, the track saw a record attendance in 2024: with a fourth information extraction task added, more runs submitted with the largest number of participating teams ever. In fact, the community is broadening beyond CLEF and raising general interest in generative scientific text summarization and simplification [30].

Acknowledgments. This research was funded, in whole or in part, by the French National Research Agency (ANR) *Automatic Simplification of Scientific Texts* project (ANR-22-CE23-0019-01) (https://anr.fr/Project-ANR-22-CE23-0019). We also thank the MaDICS research group (https://www.madics.fr/ateliers/simpletext/). The SOTA Task is jointly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number: NFDI4DataScience (460234259) and the German BMBF project SCINEXT (01lS22070).

This track would not have been possible without the great support of numerous individuals. We want to thank in particular the colleagues and the students who participated in data construction and evaluation. Please visit the SimpleText website for more details on the track (http://simpletext-project.com).

References

1. Achiam, J., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Ali, S.M., Sajid, H., Aijaz, O., Waheed, O., Alvi, F., Samad, A.: Improving scientific text comprehension: a multi-task approach with GPT-3.5 turbo and neural ranking. In: [18]
3. Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.): Working Notes of CLEF 2023: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, vol. 3497. CEUR-WS.org (2023). http://ceur-ws.org/Vol-3497
4. Bakker, J., Yüksel, G., Kamps, J.: University of Amsterdam at the CLEF 2024 simpletext track. In: [18]
5. Capari, A., Azarbonyad, H., Tsatsaronis, G., Afzal, Z.: Enhancing scientific document simplification through adaptive retrieval and generative models. In: [18]
6. Chung, H.W., et al.: Scaling instruction-finetuned language models. J. Mach. Learn. Res. **25**(70), 1–53 (2024)
7. Di Nunzio, G., Marchesin, S., Silvello, G.: A systematic review of automatic term extraction: what happened in 2022? Digit. Scholarsh. Humanit. **38**(Supplement_1), i41–i47 (2023). https://doi.org/10.1093/llc/fqad030. ISSN 2055-7671
8. Di Nunzio, G., Vezzani, F., Gallina, E.: UNIPD@SimpleText2024: a semi-manual approach on prompting ChatGPT for extracting terms and write terminological definitions. In: [18]
9. D'Souza, J., et al.: Overview of the CLEF 2024 simpletext task 4: track the state-of-the-art in scholarly publications. In: [18]
10. Elagina, R., Vučić, P.: AI contributions to simplifying scientific discourse in SimpleText 2024. In: [18]
11. Ermakova, L., Azarbonyad, H., Bertin, S., Augereau, O.: Overview of the CLEF 2023 SimpleText task 2: difficult concept identification and explanation. In: [3]. https://ceur-ws.org/Vol-3497/paper-239.pdf
12. Ermakova, L., et al.: Text simplification for scientific information access. In: Hiemstra, D., Moens, M.-F., Mothe, J., Perego, R., Potthast, M., Sebastiani, F. (eds.) ECIR 2021. LNCS, vol. 12657, pp. 583–592. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-72240-1_68
13. Ermakova, L., Bertin, S., McCombie, H., Kamps, J.: Overview of the CLEF 2023 SimpleText task 3: scientific text simplification. In: [3]. https://ceur-ws.org/Vol-3497/paper-240.pdf
14. Ermakova, L., Kamps, J.: Complexity-aware scientific literature search: searching for relevant and accessible scientific text. In: Nunzio, G.M.D., Vezzani, F., Ermakova, L., Azarbonyad, H., Kamps, J. (eds.) Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pp. 16–26, ELRA and ICCL, Torino (2024). https://aclanthology.org/2024.determit-1.2
15. Ermakova, L., SanJuan, E., Huet, S., Azarbonyad, H., Augereau, O., Kamps, J.: Overview of the CLEF 2023 SimpleText lab: automatic simplification of scientific texts. In: Arampatzis, A., et al. (eds.) CLEF 2023. LNCS, vol. 14163, pp. 482–506. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-42448-9_30
16. Ermakova, L., et al.: Overview of the CLEF 2022 SimpleText lab: automatic simplification of scientific texts. In: Barrón-Cedeño, A., et al. (eds.) CLEF 2022. LNCS, vol. 13390, pp. 470–494. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-13643-6_28
17. Ermakova, L., et al.: Overview of the CLEF 2024 SimpleText task 3: simplify scientific text. In: [18]
18. Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.): Working Notes of CLEF 2024: Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings. CEUR-WS.org (2024)
19. Jiang, A.Q., et al.: Mistral 7B. arXiv preprint arXiv:2310.06825 (2023)

20. Kabongo, S., D'Souza, J., Auer, S.: ORKG-leaderboards: a systematic workflow for mining leaderboards as a knowledge graph. arXiv preprint arXiv:2305.11068 (2023)
21. Kabongo, S., D'Souza, J., Auer, S.: Zero-shot entailment of leaderboards for empirical AI research. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2023 (2023)
22. Kabongo, S., D'Souza, J., Auer, S.: Exploring the latest LLMs for leaderboard extraction: a comprehensive investigation. In: [18]
23. Kabongo, S., D'Souza, J., Auer, S.: Automated mining of leaderboards for empirical AI research. In: Ke, H.-R., Lee, C.S., Sugiyama, K. (eds.) ICADL 2021. LNCS, vol. 13133, pp. 453–470. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-91669-5_35
24. Kardas, M., et al.: AxCell: automatic extraction of results from machine learning papers. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 8580–8594 (2020)
25. Largey, N., Maarefdoust, R., Durgin, S., Mansouri, B.: AIIR lab systems for CLEF 2024 SimpleText: large language models for text simplification. In: [18]
26. Lin, Z.: How to write effective prompts for large language models. Nat. Hum. Behav. **8**(4), 611–615 (2024). https://doi.org/10.1038/s41562-024-01847-2, https://www.nature.com/articles/s41562-024-01847-2. ISSN 2397-3374
27. Mann, R., Mikulandric, T.: CLEF 2024 SimpleText tasks 1-3: use of LLaMA-2 for text simplification. In: [18]
28. Michail, A., Andermatt, P.S., Fankhauser, T.: Scientific text simplification using multi-prompt minimum bayes risk decoding: examining MBR's decisions. In: [18]
29. Navigli, R., Velardi, P.: Learning word-class lattices for definition and hypernym extraction. In: ACL, pp. 1318–1327 (2010)
30. Nunzio, G.M.D., Vezzani, F., Ermakova, L., Azarbonyad, H., Kamps, J. (eds.): Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia (2024). https://aclanthology.org/2024.determit-1.0
31. Nunzio, G.M.D., et al.: Overview of the CLEF 2024 SimpleText task 2: identify and explain difficult concepts. In: [18]
32. OpenAI: GPT-3.5 turbo documentation (2023). https://platform.openai.com/docs/models/gpt-3-5-turbo. Accessed 10 June 2024
33. Ortiz-Zambrano, J.A., Espin-Riofrio, C., Montejo-Ráez, A.: SINAI participation in SimpleText task 2 at CLEF 2024: zero-shot prompting on GPT-4-turbo for lexical complexity prediction. In: [18]
34. SanJuan, E., Huet, S., Kamps, J., Ermakova, L.: Overview of the CLEF 2023 SimpleText task 1: passage selection for a simplified summary. In: [3]. https://ceur-ws.org/Vol-3497/paper-238.pdf
35. SanJuan, E., et al.: Overview of the CLEF 2024 SimpleText task 1: retrieve passages to include in a simplified summary. In: [18]
36. Staudinger, M., El-Ebshihy, A., Ningtyas, A.M., Piroi, F., Hanbury, A.: AMATU@Simpletext2024: are LLM alone any good for scientific entity extraction? In: [18]
37. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: ArnetMiner: extraction and mining of academic social networks. In: KDD 2008, pp. 990–998 (2008)
38. Teufel, S., et al.: Argumentative zoning: information extraction from scientific text. Ph.D. thesis, Citeseer (1999)
39. Touvron, H., et al.: Llama 2: open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
40. Varadi, D.P., Bartulović, A.: SimpleText 2024: scientific text made simpler through the use of AI. In: [18]

41. Vendeville, B., Ermakova, L., De Loor, P.: UBO NLP report on the SimpleText track at CLEF 2024. In: [18]
42. Xu, W., Callison-Burch, C., Napoles, C.: Problems in current text simplification research: new data can help. Trans. ACL **3**, 283–297 (2015). https://www.mitpressjournals.org/doi/abs/10.1162/tacl_a_00139. ISSN 2307-387X

Overview of Touché 2024: Argumentation Systems

Johannes Kiesel[1(✉)], Çağrı Çöltekin[2], Maximilian Heinrich[1], Maik Fröbe[3], Milad Alshomary[4], Bertrand De Longueville[5], Tomaž Erjavec[6], Nicolas Handke[7], Matyáš Kopp[8], Nikola Ljubešić[6], Katja Meden[6], Nailia Mirzhakhmedova[1], Vaidas Morkevičius[9], Theresa Reitis-Münstermann[10], Mario Scharfbillig[5], Nicolas Stefanovitch[5], Henning Wachsmuth[4], Martin Potthast[11,12,13], and Benno Stein[1]

[1] Bauhaus-Universität Weimar, Weimar, Germany
touche@webis.de
[2] University of Tübingen, Tübingen, Germany
[3] Friedrich-Schiller-Universität, Jena, Germany
[4] Leibniz University Hannover, Hanover, Germany
[5] European Commission, Joint Research Centre (JRC), Brussels, Belgium
[6] Jožef Stefan Institute, Ljubljana, Slovenia
[7] Leipzig University, Leipzig, Germany
[8] Charles University, Prague, Czech Republic
[9] Kaunas University of Technology, Kaunas, Lithuania
[10] Arcadia Sistemi Informativi Territoriali, Milano, Italy
[11] University of Kassel, Kassel, Germany
[12] hessian.AI, Darmstadt, Germany
[13] ScaDS.AI, Leipzig, Germany

Abstract. This paper is a condensed overview of Touché: the fifth edition of the lab on argumentation systems that was held at CLEF 2024. With the goal to foster the development of support-technologies for decision-making and opinion-forming, we organized three shared tasks: (1) Human value detection (ValueEval), where participants detect (implicit) references to human values and their attainment in text; (2) Multilingual Ideology and Power Identification in Parliamentary Debates, where participants identify from a speech the political leaning of the speaker's party and whether it was governing at the time of the speech (new task); and (3) Image retrieval or generation in order to convey the premise of an argument with visually. In this paper, we describe these tasks, their setup, and participating approaches in detail.

Keywords: Argumentation · Human values · Ideology · Image retrieval

L. Goeuriot et al. (Eds.): CLEF 2024, LNCS 14959, pp. 308–332, 2024.
https://doi.org/10.1007/978-3-031-71908-0_14

1 Introduction

Decision-making and opinion-forming are everyday tasks, for which everybody has the chance to acquire knowledge on the Web on almost every topic. However, conventional search engines are primarily optimized for returning *relevant* results, which is insufficient for collecting and weighing the pros and cons for a topic. To close this gap of technologies that support people in decision-making and opinion-forming, the Touché lab's shared tasks[1] (https://touche.webis.de) call for the research community to develop respective approaches. In 2024, we organized the three following shared tasks:

1. Human Value Detection (a continuation of ValueEval'23 @ SemEval [38]) features two subtasks in ethical argumentation of detecting human values in texts and their attainment, respectively.
2. Ideology and Power Identification in Parliamentary Debates features two subtasks in debate analysis of detecting the ideology and position of power of the speaker's party, respectively (new task).
3. Image Retrieval/Generation for Arguments (third edition, now joint task with ImageCLEF) is about the retrieval or generation of images to help convey an argument's premise.

In total, 20 teams participated in Touché in 2024. Nine teams participated in the human value detection task (cf. Sect. 4)—of which six submitted a notebook paper—and submitted 21 runs. Most teams integrated DeBERTa [32], RoBERTa [46], or the multi-lingual XLM-RoBERTa [12]. Only one team employed a generative approach (employing GPT-4o). Nine teams participated in the multilingual ideology and power identification task (cf. Sect. 5) and submitted 52 runs. The majority of teams participated in both subtasks. While traditional machine learning methods like support vector classifiers or logistic regression with n-gram features were more common among participating teams, higher-scores were typically obtained by teams using pretrained models. The two teams that participated in the image retrieval/generation task used similarity embeddings between images and text. One team used CLIP [58], the other a DPR [35] inspired approach. The corpora, topics, and judgments created at Touché are freely available to the research community on the lab's website.[2]

2 Related Work

Argumentation systems are diverse and are connected to many fields within and outside of computer science. The following sections review the related work for each Touché task of 2024.

[1] 'touché' confirms "a hit in fencing or the success or appropriateness of an argument, an accusation, or a witty point." [https://merriam-webster.com/dictionary/touche].
[2] https://touche.webis.de/.

2.1 Human Value Detection

Due to their outlined importance, human values have been studied both in the social sciences [66] and in formal argumentation [8] for decades. According to the former, a "value is a (1) belief (2) pertaining to desirable end states or modes of conduct, that (3) transcends specific situations, (4) guides selection or evaluation of behavior, people, and events, and (5) is ordered by importance relative to other values to form a system of value priorities." For cross-cultural analysis, Schwartz derived 48 value questions from universal individual and societal needs, including concepts such as *obeying all the laws* and *being humble* [67]. Based on these taxonomies are several studies in the social sciences, which could greatly benefit from the automated methods our task aims at [64]. See Scharfbillig et al. [65] for a recent overview and practical insights from the social sciences.

Moreover, several works in computer science utilize values. For example, in the context of interactive systems, to tune interactive chat-based agents or texts in general towards morally acceptable behavior [3,45]. A related dataset is ValueNet [57], which contains 21K one-sentence descriptions of social scenarios (taken from SOCIAL-CHEM-101 [23]) annotated for the 10 value categories of an earlier version of Schwartz' value taxonomy. A major difference to the Touché24-ValueEval dataset are the more ordinary situations in ValueNet (e.g., whether to say "I miss mom"). Our earlier work analyzed values in short arguments [37,38].

2.2 Ideology and Power Identification

Parliamentary data has a high societal impact and provides publicly available sources for analyzing (argumentative) language. Therefore, the number of resources based on parliamentary proceedings [22,42], and computational and linguistics analyses of parliamentary debates [1,28] increased in recent years.

The present task is about two important aspects of the political discourse, *ideology* and *power*. Although a simplification, political orientation on the left-to-right spectrum has been one of the defining properties of political ideology [5,74]. Power is another factor that shapes the political discourse [15,20,21]. Automatic identification of political orientation from texts has attracted considerable interest [10,13,27,55,56], including a few recent shared tasks [25,62]. The present task differs from the earlier ones, with respect to the source material (parliamentary debates, rather than the popular sources of social media or news) and multilinguality. Despite its central role in critical discourse analysis, to the best of our knowledge, power in parliamentary debates has not been studied computationally. There has been only a few recent computational studies providing indications of linguistic differences between governing and opposition parties [40,49,51,71]. The present shared task and associated data is likely to provide a reference for the future studies investigating power in political discourse.

2.3 Image Retrieval/Generation for Arguments

Images are a powerful tool for visual communication. They can provide contextual information and express, underline, or popularize an opinion [17], thereby taking the form of subjective statements [18]. Some images express both a premise and a conclusion, making them full arguments [30,61]. Other images may provide contextual information only and have to be combined with a textual conclusion to form a complete argument. In this regard, a recent SemEval task distinguished a total of 22 persuasion techniques in memes alone [16]. Moreover, argument quality dimensions like acceptability, credibility, emotional appeal, and sufficiency [75] all apply to arguments that include images as well.

3 Lab Overview and Statistics

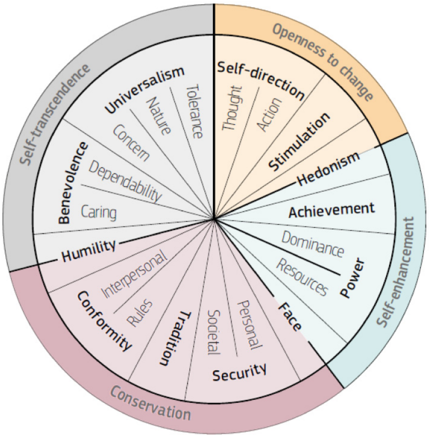
For the fifth edition of the Touché lab, we received 68 registrations from 22 countries (vs. 41 registrations in 2023). The most lab registrations came from India (24). Out of the 68 registered teams, 20 actively participated in this year's Touché edition (9, 9, and 2 teams submitting valid runs for Task 1, 2, and 3, respectively). Active teams in previous editions were: 7 in 2023, 23 in 2022, 27 in 2021, and 17 in 2020.

We used TIRA [24] as the submission platform for Touché 2024 through which participants could either submit code, software, or run files.[3] Code and software submissions increase reproducibility, as the software can later be executed on different data of the same format. To submit software, a team implemented their approach in a Docker image that they then uploaded to their dedicated Docker registry in TIRA. Software submissions in TIRA are immutable, and after the docker image had been submitted, the teams specified the to-be-executed command—the same Docker image can thus be used for multiple software submissions (e.g., by changing some parameters). A team could upload as many Docker images or software submissions as they liked; only they and TIRA had access to their dedicated Docker image registry (i.e., the images were not public while the shared task was ongoing). To improve reproducibility, TIRA executes software in a sandbox by removing the internet connection (ensuring that the software is fully installed in the Docker image which eases rerunning software later, as libraries and models must be installed in an image). For the execution, participants could select the resources that their software had available for execution, from 1 CPU core with 10 GB RAM up to 5 CPU cores with 50 GB RAM and 1 Nvidia A100 GPU with 40 GB RAM. Participants could run their software multiple times using different resources to study the scalability and reproducibility (e.g., whether the software executed on a GPU yields the same results as on a CPU). TIRA used a Kubernetes cluster with 1,620 CPU cores, 25.4 TB RAM, 24 GeForce GTX 1080 GPUs, and 4 A100 GPUs to schedule and execute the software submissions, to allocate the resources that the participants selected.

4 Task 1: Human Value Detection (ValueEval)

The goal of this task is to develop approaches that allow for the large-scale analysis of human values behind texts. In argumentation, one has to consider

[3] https://tira.io.



Inner circle: 19 human values
(see https://valueeval.webis.de)

Outer circle: four motivational directions
(not used in this task)

- **Openness to change**
  Being independent and exploring
- **Self-enhancement**
  Seeking pleasure, wealth, and esteem
- **Conservation**
  Preserving group cohesion, order, and security
- **Self-transcendence**
  Helping others, close ones, and nature

Fig. 1. The 19 values used in this task, shown in the Schwartz value taxonomy [67].

that people have different beliefs and priorities of what is generally worth striving for (e.g., personal achievements vs. humility) and how to do so (e.g., being self-directed vs. respecting traditions), referred to as (human) values. By analyzing corpora of texts, for example for news portals or political parties, one can develop an understanding of the values that the authors deem the most important.

4.1 Task Definition

The task is to identify the values of the widely accepted value taxonomy of Schwartz [67] (cf. Fig. 1) and their attainment in long texts of nine languages (Bulgarian, Dutch, English, French, German, Greek, Hebrew, Italian, and Turkish). This taxonomy has been replicated in over 200 samples in 80 countries and is the backbone of value research [65]. A value can either be mentioned as something that is or should be attained (i.e., lead towards fulfilling the value) or something that is constrained, i.e., not attained. For example, for Security, (partial) attainment would mean that something is made safer or healthier. In contrast, an event can be stated in a way that thwarts or constrains safety or health. Participating teams can submit software in one or both of two sub-tasks: (1) Given a text, for each sentence, detect which human values the sentence refers to; and (2) Given a text, for each sentence and value this sentence refers to, detect whether this reference (partially) attains or constrains the value.

4.2 Data Description

The task employs a collection of 2648 human-annotated texts in nine languages from news articles and political manifestos. Texts are sampled to reflect diverse opinions (different parties; mainstream news and others) from 2019 to 2023. The

Table 1. Overview of the Touché24-ValueEval dataset by language, with the respective number of texts, sentences, annotator agreement as measured by Krippendorf's α, and the thousandths of these sentences with any or a specific value (attained or constrained). Languages are Bulgarian (BE), German (DE), Greek (EL), English (EN), French (FR), Hebrew (HE), Italian (IT), Dutch (NL), and Turkish (TR).

				Sentences with value (‰)																			
Lang.	**Texts**	**Sentences**	α	**Any value**	**Self-direction: thought**	**Self-direction: action**	**Stimulation**	**Hedonism**	**Achievement**	**Power: dominance**	**Power: resources**	**Face**	**Security: personal**	**Security: societal**	**Tradition**	**Conformity: rules**	**Conformity: interpersonal**	**Humility**	**Benevolence: caring**	**Benevolence: dependability**	**Universalism: concern**	**Universalism: nature**	**Universalism: tolerance**
BG	260	6 919	.495	641	010	055	046	005	075	053	053	021	011	108	020	089	009	002	059	021	071	023	005
DE	261	9 183	.367	533	018	055	034	011	079	032	038	020	026	059	009	072	015	002	017	015	050	026	014
EL	328	7 349	.696	615	003	013	029	003	054	074	089	018	011	130	006	060	046	000	024	032	054	025	014
EN	408	10 305	.409	306	004	025	005	004	043	016	016	006	014	053	008	036	016	003	006	007	031	012	008
FR	219	4 650	.685	304	005	023	016	005	019	024	015	020	021	065	006	030	010	001	012	007	038	020	009
HE	250	7 331	.557	859	025	042	021	003	081	122	094	032	029	170	031	096	011	002	016	041	080	022	015
IT	276	6 379	.610	632	010	015	072	008	133	053	082	029	013	071	003	076	002	000	018	004	045	038	009
NL	323	10 982	.411	366	014	029	004	003	039	030	037	010	009	072	004	033	005	002	004	017	043	019	009
TR	323	11 133	.463	473	015	046	027	022	059	025	045	016	042	072	027	071	007	004	047	025	036	014	007
All	2 648	74 231	.546	512	012	035	026	008	063	045	050	018	020	086	013	061	013	002	022	019	048	021	010

data is annotated as part of the ValuesML project[4] by over 70 value scholars. The annotators marked segments in the texts, selected from 19 values the one that the segment refers to most, and selected whether the segment (partially) attains or constrains the value, or whether it is unclear if it attains or constrains it. Dedicated team leaders per language trained the respective annotators, consolidated annotations into a single ground truth, and discussed sentences were annotators disagreed (measured continuously by us) in their language teams. The team leaders discussed issues with us in bi-weekly meetings. Moreover, we discussed with the team leaders the current holistic inter-annotator agreement [70] and its change compared to the previous meeting to monitor annotation quality and coherence across documents and languages. To measure annotator agreement, we computed Krippendorf's α before curation for all language teams individually and overall (cf. Table 1). We see this agreement as sufficient, and belief that the curation process increased the annotation quality even further.

For Touché, the dataset is automatically split into sentences using Trankit version 1.1.1 [52] (cf. Table 2 for the sentence-based dataset format). The dataset is provided both in the original language and automatically translated to English,

[4] https://knowledge4policy.ec.europa.eu/projects-activities/valuesml-unravelling-expressed-values-media-informed-policy-making_en.

Table 2. Excerpt of the dataset for the human value detection task. The dataset comes in six directories: training, validation, and test data for both the original multilingual dataset and its automatic translation to English. Each directory contains a `sentences.tsv` where each row corresponds to one sentence. The training and validation directories also each contain a `labels.tsv` where each row corresponds to a sentence in `sentences.tsv` and columns 3–40 correspond to labels (attained and constrained for each of the 19 values). Label values in the `labels.tsv` are either 1.0 if the sentence refers to that value and attainment polarity, 0.0 if it does not, or 0.5 if the sentence refers to that value but the attainment polarity is unclear (0.2% of cases).

`sentences.tsv` (3 columns)

Text-ID	Sentence-ID	Text
EN_012	1	Who designed global guidelines for puberty blockers?
EN_012	2	More and more children and young people believe they have to question their gender...
EN_012	3	Some 60 minors were treated in the Netherlands in 2010, but has increased to around...

`labels.tsv` (40 columns)

Text-ID	Sentence-ID	Self-direction: thought attained	Self-direction: thought constrained	...
EN_012	1	0.0	0.0	...
EN_012	2	1.0	0.0	...
EN_012	3	0.0	0.0	...

either using DeepL or, for Hebrew, Google Translate.[5] The dataset is split into sets by texts, so that 60% of sentences are in the training set, 20% in the validation set, and 20% in the test set.[6]

Table 1 shows the size of the dataset for each language and the value distribution. The number of texts per language are between 219 (French) and 408 (English). The number of sentences per language are between 4 650 (French) and 11 133 (Turkish). Only 30.4% of the French sentences are annotated as referring to a value, but 85.9% of Hebrew sentences. The least frequent value overall is *Humility* (0.2%) and the most frequent one is *Security: societal* (8.6%). This in-balance between languages and values makes the multi-label classification problem especially challenging.

4.3 Participant Approaches

In 2024, nine teams participated in this task (of which six submitted a notebook paper) and submitted 21 runs. Moreover, we added two baseline runs for comparison. Five of the six teams that submitted a paper relied on DeBERTa [32], RoBERTa [46], or the multi-lingual XLM-RoBERTa [12]. The other team (Eric Fromm) used GPT-4o.[7] Two teams work with the multi-lingual dataset (Arthur Schopenhauer, Hierocles of Alexandria) whereas the others use the English trans-

[5] https://www.deepl.com/pro-api and https://cloud.google.com/translate.
[6] Dataset: https://zenodo.org/doi/10.5281/zenodo.10396293.
[7] https://openai.com/index/hello-gpt-4o/.

lations only. Only one team (Hierocles of Alexandria) used the sentence sequence, whereas the other teams classified each sentence individually.

Baselines. We provide two baselines, that also served to kickstart the participants' approaches:[8] (1) a random baseline that assigns a (uniformly) random value "confidence" to each value for each sentence in subtask 1 and randomly distributes this confidence between attained and constrained for subtask 2; and (2) a BERT [14] baseline with a multi-label classification head for all 38 combinations of value and attainment.

Team Arthur Schopenhauer [77].[9] The team used the multi-lingual dataset and analyzed the sentences independently. They approached subtask 1 as a classification problem. A *no-label* class was added for sentences without assigned value, and sentences with *Humility* were ignored due to the scarcity of that value. The 6% of sentences with more than one assigned value were ignored, as well. Different models were fine-tuned for English texts (deberta-v2-xxlarge [32]) and others (xlm-roberta-large [12]). In both cases, an ensemble with a thresholded soft voting scheme of four models was employed: one model for each combination of two seeds and two loss functions. For loss functions the authors report that cross entropy lead to higher results in their preliminary tests for frequent values but weighted cross entropy did so for infrequent values. The team approached subtask 2 as a binary classification problem, ignoring the few sentences with *unknown* attainment. Their approach is otherwise the same as for subtask 1, except that only a single model was employed instead of an ensemble (with cross entropy loss) based on results from their preliminary tests.

Team Edward Said [7]. The team used the English translations of the dataset and analyzed the sentences independently. To counter the label imbalance, the team upsampled sentences by a factor of four if the associated label is one of 14 underrepresented labels (value + attainment). They selected these 14 labels out of the 38 labels if the label was infrequent in total or in comparison to the other label for the same value (but different attainment). They then fine-tuned a RoBERTa [46] and DeBERTa [32] model for multi-label classification.

Team Eric Fromm [50]. The team used the English translations of the dataset and analyzed the sentences independently. They employed GPT-4o for zero-shot classification, prompting it with the 19 value descriptions from the annotator's guide to select one or none for each sentence. They did not tackle subtask 2.

Team Hierocles of Alexandria [41].[10] The team used both the multi-lingual dataset and English translations and incorporated sentence sequence information. More specifically, their approach predicts values for a sentence from an

[8] https://github.com/touche-webis-de/touche-code/tree/main/clef24/human-value-detection/approaches.

[9] Code: https://github.com/h-uns/clef2024-human-value-detection.

[10] Code: https://github.com/SotirisLegkas/Touche-ValueEval24-Hierocles-of-Alexandria.

input text that consists of the previous two sentences concatenated with the target sentence. The two preceding sentences contained special tokens to represent any values assigned to them. During training and validation the true labels were employed, but during testing the predicted labels of the previous sentences were leveraged. The team fine-tuned different RoBERTa [46] and DeBERTa [32] models for English and XLM-RoBERTa [12] models for the multi-lingual dataset, with the best performing one being XLM-RoBERTa-xl [29]. Moreover, they developed a custom model architecture for multi-label text classification consisting of multiple classification heads. Each classification head focused on a different language for the multi-lingual dataset. The custom model architecture was adapted and employed for the English-translated dataset as well. After preliminary experiments concerning loss functions, class weights and various thresholds, they used the binary cross-entropy loss with logits as their loss function and selected an optimal classification threshold for each value. The approach is trained to tackle both subtasks 1 and 2.

Team Philo of Alexandria [76].[11] The team used the English translations of the dataset and analyzed the sentences independently. They approached subtask 1 as a multi-label problem and fine-tuned DeBERTa (deberta-base [32]) after initial experiments with several models. They employ the same base model for subtask 2 and fine-tune it to classify each text pair of sentence and human value name into either attaining or constraining.

Team SCaLAR NITK (code name: Peter Abelard) [34]. The team used the English translations of the dataset and analyzed the sentences independently. They experimented with SVMs, KNNs, decision trees, hierarchical classification, transformer models and large language models. Based on preliminary experiments, they fine-tuned a RoBERTa [46] model for both subtasks (multi-label and binary classification, respectively).

4.4 Task Evaluation

Following ValueEval'23 [38], submissions are evaluated using standard macro F_1-score over all values. The same metric is used for the new subtask 2. The submission format has been designed so that participants submit only one run file for both subtasks (same format as the `labels.tsv`), but the scores for the subtasks are calculated independently of each other from the same file as follows. Each submission includes for each sentence and value a confidence score (between 0 and 1) for both attained and constrained polarity. If the sum of the two numbers is above 0.5, the submission is evaluated as having predicted that the sentence refers to that value (subtask 1). For subtask 2, only the sentence-value pairs are considered for which the sentence refers to the value according

[11] Code: https://github.com/VictorMYeste/touche-human-value-detection
Models: https://huggingface.co/VictorYeste/deberta-based-human-value-detection
https://huggingface.co/VictorYeste/deberta-based-human-value-stance-detection
Image: `docker pull victoryeste/valueeval24-philo-of-alexandria-deberta-cascading`.

to the ground-truth. For these pairs, the submission is evaluated as having predicted the attainment polarity for which it produced the larger confidence score.

Table 3 shows the results for the best-performing approaches per team for both subtasks. The best-performing approach for subtask 1 is the one of team Hierocles of Alexandria that uses XLM-RoBERTa-xl, the previous sentences, and is trained specifically for subtask 1. Overall, multilingual models performed best, with also the second-in-place employing such a model. Rarer values are overall detected worse, with the exception of the zero-shot approach by team Eric Fromm (especially Humility), indicating insufficient training data. Several teams achieved top scores for subtask 2. Overall, this binary classification task is, as once can expect, much easier than subtask 1. However, most teams clearly focused their efforts on subtask 1, so there is likely more room for improvement.

5 Task 2: Multilingual Ideology and Power Identification in Parliamentary Debates

The study of parliamentary debates is crucial to understand the decision processes in the parliaments and their societal impacts. The goal of this task is to automatically identify two important aspects of parliamentary debates: the political orientation of the party of the speaker, and the role of the party of the speaker in the governance of the country or the region. Identifying these underlying aspects of parliamentary debates enables automated comprehension of these discussions, the decisions that these discussions lead to, and their consequences.

5.1 Task Definition

Both subtasks were defined as binary classification tasks: Given a parliamentary speech, (1) predict the political orientation of the party of the speaker on the *left–right* spectrum, and (2) predict whether the speaker belongs to one of the governing parties or the opposition. The first task is relatively well studied, and there have been some recent shared tasks on identifying political orientation [25,62]. Unlike the earlier tasks, our data set includes multiple parliaments and languages, and is based on parliamentary debates. To the best of our knowledge, automatic identification of governing role—power—has not been studied earlier.

5.2 Data Description

The source of the data for this task is the ParlaMint [19], a uniformly encoded and annotated corpus of transcripts of parliamentary speeches from multiple national and regional parliaments.[12] The transcripts are The ParlaMint version 4.0 used for the task includes data from the following national and regional

[12] Although all transcripts are obtained thorough the data published by the respective parliaments, the method for obtaining the transcripts vary, such as scraping the web site of the parliament, extracting from published PDF files, and obtaining through an API provided by the parliament. For details, we refer to [19].

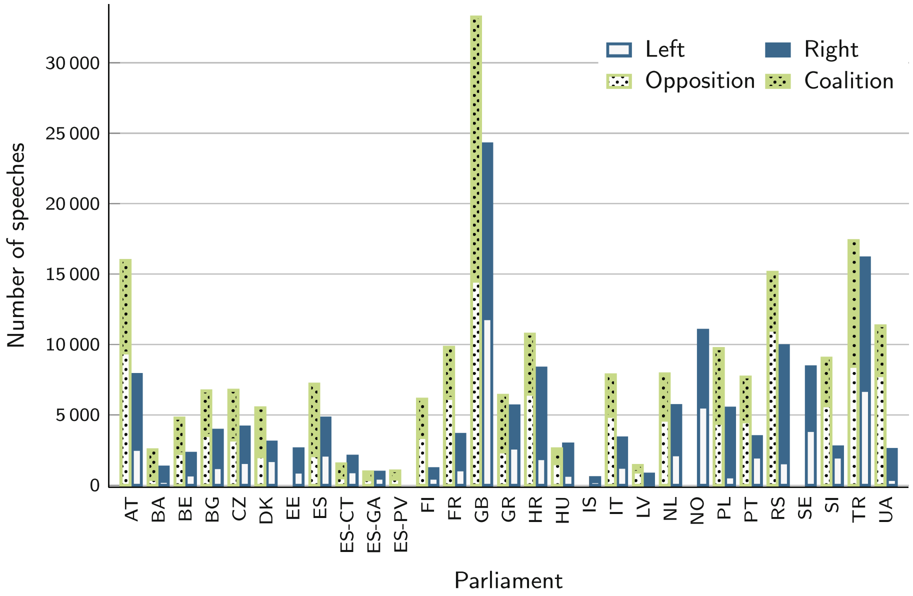


Fig. 2. Overview of the Touché24 ideology and power identification dataset. The bars show the training set for both subtasks for each parliament. Test set sizes are approximately 2 000 speeches for all parliaments.

parliaments: Austria (AT), Bosnia and Herzegovina (BA), Belgium (BE), Bulgaria (BG), Czechia (CZ), Denmark (DK), Estonia (EE), Spain (ES), Catalonia (ES-CT), Galicia (ES-GA), Basque Country (ES-PV), Finland (FI), France (FR), Great Britain (GB), Greece (GR), Croatia (HR), Hungary (HU), Iceland (IS), Italy (IT), Latvia (LV), The Netherlands (NL), Norway (NO), Poland (PL), Portugal (PT), Serbia (RS), Sweden (SE), Slovenia (SI), Turkey (TR) and Ukraine (UA). The labels for both subtasks are also coded in the ParlaMint corpora. For the sake of simplicity, we formulate both tasks as binary classification tasks. For both tasks, the main challenge in the creation of a dataset is to minimize the effects of covariates. Even though the instances to classify are speeches, the annotations are based on the party membership of the speaker. As a result, underlying variables like party membership, or speaker identity perfectly covary with ideology and power in most cases.

As a trade-off between data size, and for reducing the effect of covariates, we opt for a speaker-based sampling. First, to discourage, to some extent, the classifiers from relying on author identification, we sample at most 20 speeches of a single speaker. This is also important for introducing variation into the dataset, as the number of speeches from each speaker follows a power-law distribution: While a small number of speakers tend to deliver most of the speeches, e.g., party or party group leaders, most speakers have relatively few speeches. The distribution of speeches or speakers to include in training and test sets is also

Table 3. Achieved F_1-score of the best submission per team (as measured by overall F_1-score) on the test dataset for subtasks 1 and 2, and whether the submission used the original multilingual dataset or the automatic translation to English (EN). Baseline submissions ("Aristotle") are shown in gray.

Subtask 1

		F_1-score																			
Team	**Lang.**	**Overall**	**Self-direction: thought**	**Self-direction: action**	**Stimulation**	**Hedonism**	**Achievement**	**Power: dominance**	**Power: resources**	**Face**	**Security: personal**	**Security: societal**	**Tradition**	**Conformity: rules**	**Conformity: interpersonal**	**Humility**	**Benevolence: caring**	**Benevolence: dependability**	**Universalism: concern**	**Universalism: nature**	**Universalism: tolerance**
Hierocles of Alexandria [41]	multil.	39	15	27	30	37	45	42	49	31	42	49	46	51	24	00	34	33	47	63	27
Arthur Schopenhauer [77]	multil.	35	12	24	33	35	40	37	47	24	38	46	49	50	19	00	32	31	46	60	27
Philo of Alexandria [76]	EN	28	08	22	27	31	35	31	34	17	33	40	47	42	09	00	21	28	40	57	21
SCaLAR NITK [34]	EN	28	05	17	27	27	38	34	38	15	34	40	41	43	07	00	23	26	37	56	16
Edward Said [7]	EN	28	05	17	11	15	25	31	34	16	32	41	45	44	06	05	10	23	41	57	27
Erich Fromm [50]	EN	25	15	10	10	18	25	18	09	24	21	30	46	33	09	15	26	15	41	55	20
Lawrence Kohlberg	EN	25	08	11	19	23	31	22	31	11	28	37	34	42	09	00	21	23	34	54	18
Aristotle (BERT)	EN	24	00	13	24	16	32	27	35	08	24	40	46	42	00	00	18	22	37	55	02
John Shelby Spong	EN	07	00	00	02	00	16	05	11	00	01	28	00	15	00	00	00	00	13	27	00
Alain Badiou	EN	07	00	00	02	00	16	05	11	00	01	28	00	15	00	00	00	00	13	27	00
Aristotle (random)	EN	06	02	07	05	02	11	08	10	03	04	14	03	11	03	00	05	04	09	04	02

Subtask 2

Team	**Lang.**	**Overall**	**Self-direction: thought**	**Self-direction: action**	**Stimulation**	**Hedonism**	**Achievement**	**Power: dominance**	**Power: resources**	**Face**	**Security: personal**	**Security: societal**	**Tradition**	**Conformity: rules**	**Conformity: interpersonal**	**Humility**	**Benevolence: caring**	**Benevolence: dependability**	**Universalism: concern**	**Universalism: nature**	**Universalism: tolerance**
Arthur Schopenhauer [77]	multil.	83	77	83	85	88	87	73	84	80	82	84	78	80	79	74	91	89	86	85	81
Edward Said [7]	EN	83	77	82	85	88	88	79	80	77	84	84	85	80	80	76	90	86	85	85	78
Philo of Alexandria [76]	EN	82	85	80	85	91	86	79	80	78	85	80	82	77	78	77	93	89	84	83	79
Aristotle (BERT)	EN	81	83	79	86	88	84	77	80	74	84	81	78	78	79	87	89	86	85	81	78
John Shelby Spong	EN	81	81	77	83	88	88	77	79	76	83	82	85	76	81	84	90	85	81	81	79
Alain Badiou	EN	81	81	77	83	88	88	77	79	76	83	82	85	76	81	84	90	85	81	81	79
Hierocles of Alexandria [41]	multil.	77	73	73	77	75	78	77	79	71	78	79	77	78	74	25	74	77	78	84	71
SCaLAR NITK [34]	EN	77	69	72	78	73	79	77	79	71	78	81	79	77	70	70	77	76	79	80	71
Erich Fromm [50]	EN	70	71	69	73	70	72	74	73	67	60	66	76	70	68	73	75	71	70	73	67
Lawrence Kohlberg	EN	66	81	77	83	80	70	76	63	56	33	45	85	63	46	84	90	79	69	70	60
Aristotle (random)	EN	52	51	47	54	52	53	55	53	52	52	50	54	53	49	45	53	56	52	49	56

important for proper evaluation. For the ideology task, the set of speakers in the training and test sets are disjoint. The ideal dataset split for the power identification task requires a different constraint: training and test sets should include speeches from the same speaker with different power roles. To come as close as possible to this ideal split, we opt for a best-effort training–test split.

When possible, we make sure that the speakers in the test set are also available in the training set with the opposite power role. Otherwise, we randomly sample more speakers to obtain the test set.

For evaluation, we set the test set size to 2 000 instances for both subtasks (100 to 200 speakers depending on the individual corpus and the task). Despite multiple speeches from each speaker, due to missing annotations and the lack of diversity of orientation in some parliaments, the disjoint speakers constraint mentioned above results in a small number of instances in the training set for some of the parliaments. Not all parliamentary data provides both labels. Some countries do not have the opposition–governing party distinction, and for the Galician parliament, the number and distribution of orientation labels did not result in a test set that was large enough. Figure 2 shows the training set sizes for each parliament. The test set size for all parliaments is approximately 2000 speeches. We do not provide a validation set. We provide further details on the data set and the sampling procedure in a separate publication [11].[13]

In addition ot the original speech transcripts and labels, we also provide automatic English translations, an anonymized speaker ID and the speaker's sex in the data for both tasks. Except the speaker ID, which is not in the test sets.

Both data sets exhibit a mild class and text length imbalance between parliaments. The data set's size was a technical challenge for some participants. The average text length is approximately 600 space-separated tokens, which is larger than the maximum accepted by many of the pretrained language models. Moreover, the data set is also large overall (more than 3GB uncompressed).

5.3 Participant Approaches

In 2024, 9 teams participated in this task and submitted 52 runs. We added a baseline for comparison. Unlike the ValueEval task, where pretrained language models were the dominant classifiers, for this task many participants preferred traditional, 'computationally light' approaches. A possible reason may be the large text size which is more costly to process with larger systems. Most teams, even the teams that used language models with large context sizes, truncated the texts to alleviate computational requirements. Some of the interesting improvements include ensemble of classifiers, data augmentation through back-translation and synonym replacement, multi-task learning, additional features, such as sentiment scores, and the use of domain-specific models.

Baselines. We provided only a single logistic regression baseline with tf-idf weighted character n-grams. The baseline is intentionally kept simple to encourage participation by early researchers, and reduce the computation requirements.

Team Policy Parsing Panthers [54]. The team did a set of experiments with original transcripts and their English translations, using various deep pretrained models, including BERT [14], mBERT [14], RoBERTa [46], XLM-RoBERTa [12],

[13] Training and test data are available at https://zenodo.org/doi/10.5281/zenodo.10450640, and https://zenodo.org/doi/10.5281/zenodo.11061649 respectively.

DeBERTa-v3 [32] Gemma [47] and ensembles of these models. This team presents an extensive set of approaches, and their analyses. A few interesting approaches worth mentioning in this short summary includes (1) Data augmentation and balancing through back-translation, (2) experiments with additional metadata, (3) multi-task learning, (4) the use of automatically obtained polarity labels, and increasing the number of instances in the training set of the orientation subtask by using the matching speaker IDs in the power dataset. This team participated in both subtasks for all parliaments.

Team Trojan Horses [48]. The team experimented with improving the logistic regression baseline, as well as fine-tuning BERT. They used the English translations and participated in both subtasks for the majority of the parliaments.

Team Pixel Phantoms [31]. The team experimented with some of the traditional classifiers (SVMs, logistic regression and decision trees) using the English translations provided. As well as tf-idf weighted features, they also extracted text embeddings from DistilBERT [63], through Sentence BERT [60]. They participated in both subtasks for the majority of the parliaments.

Team Ssnites [73]. The team fine-tuned BERT for the majority of parliaments and both subtasks. They relied on the English translations provided, and participated in both subtasks for the majority of the parliaments.

Team Hale Lab [68]. After some initial experiments with BERT, the team used a variety of classification methods including simple feed-forward networks, and LSTMs. The features for the models were either bag-of-words features weighted with tf-idf, or the multilingual LASER [6] embeddings. They used the original (untranslated) data, using various libraries for tokenization and preprocessing, and participated in both subtasks for the majority of the parliaments.

Team Vayam Solve Kurmaha [69]. This team also experimented with multiple traditional classification methods (SVM, kNN, random forests) and their ensembles, using the English translations. The team also used data augmentation through synonym replacement. They participated in both subtasks for the majority of the parliaments.

Team Gerber [26]. The team used a convolutional neural network (CNN) for the task without any pretrained embeddings. They used the original transcripts only, and participated in both subtasks for the majority of the parliaments.

Team JU_NLP_DID [36]. The team used SVM classifiers with tf-idf features, participating in both subtasks for the majority of the parliaments. They also make use of automatic sentiment labels as an additional feature.

Team INSA Passau [4]. The team also experimented with multiple approaches, where some of their submissions were focused on orientation identification and a smaller number of parliaments. The methods used included training SVMs, fine-tuning BERT-based models (pre)trained on legal documents [9,79] and fine-tuning and zero- and few-shot prompting the Llama [72] version 3 models with varying sizes (which were released during while the shared task was running).

Table 4. F_1-scores of the best submissions per team (as measured by overall F_1-score) on ideology identification task. Baseline scores are shown in gray.

	F_1-score																												
Team	**Overall**	**AT**	**BA**	**BE**	**BG**	**CZ**	**DK**	**EE**	**ES**	**ES-CT**	**ES-GA**	**FI**	**FR**	**GB**	**GR**	**HR**	**HU**	**IS**	**IT**	**LV**	**NL**	**NO**	**PL**	**PT**	**RS**	**SE**	**SI**	**TR**	**UA**
Policy Parsing Panthers	79	77	51	71	77	63	84	64	94	80	98	77	75	92	89	65	87	71	77	67	71	82	88	95	79	95	78	93	83
gerber	63	60	45	54	62	52	56	00	77	66	76	54	58	76	72	51	69	00	60	49	59	00	72	69	64	00	58	84	73
HALE Lab	61	56	44	59	60	52	56	52	76	69	84	52	48	74	71	43	67	57	60	49	53	61	62	67	55	77	49	83	60
Pixel Phantoms	59	58	49	56	56	47	56	54	72	64	75	59	58	72	71	55	68	57	57	54	60	54	59	54	51	61	47	78	56
Ssnites	59	50	53	55	53	50	61	52	61	58	64	55	56	64	59	53	60	58	53	51	56	66	71	64	64	75	58	79	53
Trojan Horses	59	61	25	57	61	51	60	57	72	67	00	33	60	73	74	53	71	55	66	00	60	61	68	63	00	74	00	80	68
INSA Passau	59	60	53	54	61	47	57	53	63	61	66	34	58	69	59	56	66	56	56	54	56	58	69	55	61	66	51	80	62
JU_NLP_DID	57	53	42	42	55	51	60	57	69	57	70	00	50	71	63	43	60	55	61	47	56	59	51	67	48	73	46	77	57
Baseline	56	52	42	45	53	52	56	47	72	65	67	54	43	74	74	43	57	39	56	45	51	62	46	63	53	75	39	84	58

5.4 Task Evaluation

We use macro-averaged F_1-score as the main evaluation metric for both subtasks. Similar to the ValueEval task, the participants were encouraged to submit confidence scores, where a score over 0.5 is interpreted as class 1 and otherwise 0.

Table 4 and Table 5 present the overall best-performing approaches per team for the ideology and power subtasks respectively. The best scores for both tasks are from the team Policy Parsing Panthers. The team used an ensemble of multiple models, with multiple improvements including data augmentation and multi-task learning. Results on the tables do not include approaches that were focused on only one or a small number of parliaments. A noteworthy focused submission for only GB and ideology subtask by the team INSA Passau based on fine-tuning the most recent Llama 3 model achieved the second-best result for this parliament. Although the results on both tasks are higher than the baseline we provided, the variation in the scores indicate that there is quite some room for improvement for each of the approaches.

We also observe that, as formulated in this task, identifying orientation is slightly more difficult than identifying power. The overall success of the systems on a particular parliament depends on, among others, size and class distribution of the training data, and composition of the parliament. For example, we observe a general trend (with some exceptions) that for parliaments with few or no government and opposition role changes in the data (e.g., HU, PL, and TR) the roles are easier to predict than for parliaments with more varied composition and more role changes (e.g., AT, BA, and UA).

Table 5. F_1-scores of the best submissions per team (as measured by overall F_1-score) on power identification task for each parliament. Baseline scores are shown in gray.

	F_1-score																									
Team	**Overall**	**AT**	**BA**	**BE**	**BG**	**CZ**	**DK**	**ES**	**ES-CT**	**ES-GA**	**ES-PV**	**FI**	**FR**	**GB**	**GR**	**HR**	**HU**	**IT**	**LV**	**NL**	**PL**	**PT**	**RS**	**SI**	**TR**	**UA**
Policy Parsing Panthers	83	88	56	74	81	78	87	88	91	98	90	80	82	83	95	75	97	78	75	74	90	85	84	81	94	65
HALE Lab	70	69	46	61	68	69	70	65	85	88	78	65	67	75	82	68	88	69	62	64	78	65	69	61	84	49
Trojan Horses	69	72	57	63	67	63	68	69	82	85	74	39	66	72	83	67	86	72	64	64	74	65	75	62	83	56
gerber	68	68	51	60	66	64	63	72	80	86	74	60	71	72	68	63	87	52	63	64	77	66	73	58	84	48
Vayam Solve Kurmaha	68	48	48	65	69	68	69	72	83	87	76	35	66	47	85	67	88	72	62	68	75	67	75	63	85	48
Pixel Phantoms	66	70	50	59	63	65	69	65	64	77	69	61	64	73	72	57	80	69	58	62	70	66	69	60	80	52
Baseline	64	66	45	61	68	64	56	65	78	83	71	56	66	71	63	60	86	43	51	62	76	62	65	53	83	46
JU_NLP_DID	63	68	47	55	58	57	67	60	78	55	72	00	59	00	77	65	83	71	47	63	70	63	54	56	78	43
INSA Passau	62	67	45	60	66	65	54	65	00	00	00	56	66	72	56	61	85	45	52	64	77	62	63	54	84	47
Ssnites	60	66	45	58	60	61	61	62	58	62	60	60	65	60	69	65	79	62	54	57	62	58	60	57	61	46

6 Task 3: Image Retrieval/Generation for Arguments (joint Task with ImageCLEF)

Images provide powerful visual communication, are usually perceived before text is read, and can appeal directly to our emotions. The goal of this task is to find images that convey premises. The proper use of an image can increase the persuasiveness of an argument. In this regard, images can increase the pathos [59], which is the effect an argument has on its audience.

6.1 Task Definition

This observation leads to our task, in which participants are asked to find images based on an argument that help to convey the premise of the argument. In this context, "convey" is meant in broad terms; it can represent what is described in the argument, but it can also show a generalization (e.g., a symbolic image that illustrates a related abstract concept) or a specialization (e.g., a concrete example). There is a difference between verbal language and images. Verbal language provides clear but limited information, while images provide more information than written words, but are not as precise [39]. Therefore, images alone can be ambiguous and difficult to understand without context, e.g. when they refer to symbolism. For this reason, we offer the option of submitting a rationale together with the image. The rationale is an explanatory statement that assists in understanding the picture. For example, it can be a caption or contextual information about the image. The image and the rationale are evaluated together to see how this combination conveys the premise. Participants can choose to use a retrieval approach, where they submit images from a provided dataset, or a generation-based approach, where suitable images can be generated using a model of their choice. In each submission, a participant can submit up to 10 images in a ranking order for an argument.

```
<argument>
  <id>36062-a-3</id>
  <topic>Should boxing be banned?</topic>
  <premise>
    The idea of winning through intentional infliction of pain and harm
    to another person can nurture a violent and destructive mentality.
  </premise>
  <claim>
    Boxing poses both physical and psychological threats to
    participants, hence it should be banned.
  </claim>
  <stance>pro</stance>
  <type>ANECDOTAL</type>
</argument>
```

Fig. 3. Example argument from the data set. The argument consists of an id, a premise and a claim. We also indicate the topic of the argument, as well as the argument's stance on the topic. The type element indicates that the arguments relies on anecdotal evidence. Only arguments of this type are used in our dataset.

6.2 Data Description

For the task we prepared a dataset[14] containing 136 arguments and over 9000 images. The arguments were generated with GPT-4 [2] and correspond to 24 topics. The topics were taken from various IBM datasets[15] and previous Touché Shared Tasks[16]. Each generated argument consists of a premise and a claim, and can take a pro or con stance on the topic. An example of an argument can be seen in Fig. 3. Each of the images in the dataset is tagged with additional information, such as the URL and content of the corresponding website. In addition, we have provided an analysis of each image using the Google Cloud Vision API, as well as an automatically generated caption using LLaVA [44].

6.3 Participant Approaches

In 2024, 2 teams participated in this task and submitted 8 runs. All teams chose the retrieval-approach. Moreover, we added 2 baseline runs for comparison.

Baselines. The first baseline is BM25, where the corresponding documents are the image captions from the data set and the query is the premise of the argument. In the second baseline, keywords are first extracted from the image captions. Then embeddings for the premise of an argument and the keywords are generated with SBERT [60]. A corresponding relevance score is calculated based on the cosine similarity between the embeddings and averaging them. The most relevant images are selected for submission.

[14] https://zenodo.org/records/11045831.
[15] https://research.ibm.com/haifa/dept/vst/debating_data.shtml.
[16] https://touche.webis.de/shared-tasks.html.

DS@GT [53]. The team uses CLIP [58] to embed each argument and each image in a common embedding space. The first approach ranks images by cosine similarity of the embeddings. The second approach compares for each argument the 40 highest ranked images to images that are generated to support or attack the argument. The most similar images are submitted.

HTW-DIL [33]. The team has chosen an approach inspired by DPR [35]. It applies a fine-tuned multimodal Moondream model based on the Phi 1.5 LLM [43] and uses SigLIP [78] for its vision capabilities. To generate synthetic training data, the team uses GPT-4 to generate arguments from the available image/web page data. Combinations of positive and negative argument-image pairs are used for training. The results are obtained by maximising the cosine similarity for argument and image embeddings.

6.4 Task Evaluation

For each argument and each submission, the best 5 images together with the rationales are evaluated by a human expert. This expert knows neither the rank of the image nor the team that submitted it. To facilitate the annotation, we prepared a narrative for each argument that describes what a conveying image should generally show. Therefore, each combination of image, argument and rationale is rated on a three-point Likert scale from 0 to 2, where 0 means that the image does not convey the premise at all, 1 stands for partial conveyance and 2 means that the image conveys the premise completely. For seven topics, only very few relevant images could be submitted by the participating teams, so we removed these topics, resulting in a total number of 104 arguments for the evaluation. For each submission, we first calculated the NDCG score for each argument. For the required IDCG, we have considered all submitted image, argument and justification triples submitted for the corresponding argument. The final score of a submission is the average of all NDCG scores for all arguments. The results of the shared task can be seen in Table 6. To conclude, it can be said that the relevance of an image is often determined by implicit assumptions and is subject to interpretation. Therefore, the identification of conveying images is still a very challenging task.

7 Conclusion

The fifth edition of the Touché lab on argumentation systems featured three tasks: (1) Human Value Detection, (2) Ideology and Power Identification in Parliamentary Debates, and (3) Image Retrieval/Generation for Arguments. In contrast to previous years, the focus this year was more on classification than retrieval tasks. Furthermore, two of the three tasks were multilingual, although automatic English transcriptions were provided to facilitate participation. We expanded the scope of Touché with the new tasks on human values and political power and orientation. In addition, we methodically extended the retrieval task

Table 6. NDCG values for the top 5, top 3, and most relevant image(s). The approachesare sorted according to the NDCG@5 score.

Rank	Team	Approach	NDCG@5	NDCG@3	NDCG@1
1	HTW-DIL	Ada-Summary	0.428	0.409	0.404
2	HTW-DIL	Moondream-Text	0.363	0.355	0.356
3	HTW-DIL	Moondream-Default-Image-Text	0.293	0.302	0.317
4	Baseline	BM25	0.284	0.273	0.293
5	Baseline	SBERT	0.232	0.225	0.221
6	DS@GT	Generated-Image-Clip	0.180	0.178	0.197
7	HTW-DIL	Moondream-Image-Text-EP3	0.150	0.163	0.183
8	HTW-DIL	Moondream-Image	0.146	0.155	0.178
9	DS@GT	Base-Clip-Submission	0.123	0.111	0.106
10	HTW-DIL	Moondream-Image-Text	0.120	0.140	0.178

by allowing participants to generate images instead of retrieving them. Unfortunately, no team submitted generated images in the end.

Of the 68 registered teams, 20 participated in the tasks and submitted a total of 81 runs. Participants mainly used classification architectures, with BERT and variants still very dominant, although more classical machine learning models were also used in the Ideology and Power Identification in Parliamentary Debates task. Generative models, on the other hand, were rarely used. Although the Image Retrieval/Generation for Arguments task changed to seeking images for a specific argument rather than a topic, the approaches submitted were similar to previous years. They embedded the images from the collection and then used the similarity to the query for ranking, either by embedding the query directly or generating images for the query and embedding those.

We plan to continue Touché as a collaborative platform for researchers in argumentation systems. All Touché resources are freely available, including topics, manual relevance, argument quality, and stance judgments, and submitted runs from participating teams. These resources and other events such as workshops will help to further foster the community working on argumentation systems.

Acknowledgments. This work was partially supported by the European Commission under grant agreement GA 101070014 (https://openwebsearch.eu) and the German Research Foundation under project 455911521 (LARGA) as part of the SPP 1999 (RATIO). The ideology and power identification shared task has been supported by CLARIN ERIC, under the ParlaMint project (https://www.clarin.eu/parlamint).

References

1. Abercrombie, G., Batista-Navarro, R.: Sentiment and position-taking analysis of parliamentary debates: a systematic literature review. J. Comput. Soc. Sci. **3**(1), 245–270 (2020)
2. Achiam, J., et al.: GPT-4 Technical Report (2024)
3. Ammanabrolu, P., Jiang, L., Sap, M., Hajishirzi, H., Choi, Y.: Aligning to social norms and values in interactive narratives. In: Carpuat, M., de Marneffe, M., Ruíz, I.V.M. (eds.) Proceedings of NAACL-HLT 2022, pp. 5994–6017. ACL (2022). https://doi.org/10.18653/v1/2022.naacl-main.439
4. Andruszak, M., Alhamzeh, A., Egyed-Zsigmond, E., Carlsson, A., Leydet, J., Otiefy, Y.: Team INSA Passau at Touché: multi-lingual parliamentary speech classification. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
5. Arian, A., Shamir, M.: The primarily political functions of the left-right continuum. Comp. Polit. **15**(2), 139–158 (1983)
6. Artetxe, M., Schwenk, H.: Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Trans. Assoc. Comput. Linguist. **7**, 597–610 (2019). https://doi.org/10.1162/tacl_a_00288
7. Aydin, A., Shaar, S., Cardie, C.: Edward said at touché: human values classification. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
8. Bench-Capon, T.: Persuasion in practical argument using value-based argumentation frameworks. J. Logic Comput. **13**(3), 429–448 (2003). https://doi.org/10.1093/logcom/13.3.429
9. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I.: LEGAL-BERT: the muppets straight out of law school. In: Cohn, T., He, Y., Liu, Y. (eds.) Findings of ACL: EMNLP 2020, pp. 2898–2904. ACL (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.261
10. Chen, C., Walker, D., Saligrama, V.: Ideology prediction from scarce and biased supervision: learn to disregard the "what" and focus on the "how"! In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of ACL (Volume 1: Long Papers), Toronto, Canada, pp. 9529–9549. ACL (2023). https://doi.org/10.18653/v1/2023.acl-long.530
11. Çöltekin, Ç., Kopp, M., Katja, M., Morkevicius, V., Ljubešić, N., Erjavec, T.: Multilingual power and ideology identification in the parliament: a reference dataset and simple baselines. In: Fiser, D., Eskevich, M., Bordon, D. (eds.) 4th Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora, pp. 94–100. ELRA and ICCL (2024). https://aclanthology.org/2024.parlaclarin-1.14
12. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (eds.) Proceedings of ACL, pp. 8440–8451. ACL (2020). https://doi.org/10.18653/v1/2020.acl-main.747
13. Conover, M.D., Gonçalves, B., Ratkiewicz, J., Flammini, A., Menczer, F.: Predicting the political alignment of Twitter users. In: Proceedings of PASSAT and SocialCom, pp. 192–199. IEEE (2011). https://doi.org/10.1109/PASSAT/SocialCom.2011.34
14. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio,

T. (eds.) Proceedings of NAACL-HLT, pp. 4171–4186. ACL (2019). https://doi.org/10.18653/V1/N19-1423
15. van Dijk, T.: Discourse and Power. Bloomsbury Publishing (2008)
16. Dimitrov, D., et al.: SemEval-2021 task 6: detection of persuasion techniques in texts and images. In: Proceedings of SemEval, pp. 70–98. ACL (2021). https://doi.org/10.18653/v1/2021.semeval-1.7. https://aclanthology.org/2021.semeval-1.7
17. Dove, I.J.: On images as evidence and arguments. In: van Eemeren, F., Garssen, B. (eds.) Topical Themes in Argumentation Theory, vol. 22, pp. 223–238. Springer, Dordrecht (2012). https://doi.org/10.1007/978-94-007-4041-9_15
18. Dunaway, F.: Images, emotions, politics. Mod. Am. Hist. **1**(3), 369–376 (2018). https://doi.org/10.1017/mah.2018.17
19. Erjavec, T., Ogrodniczuk, M., et al.: The ParlaMint corpora of parliamentary proceedings. LREC **57**, 415–448 (2022). https://doi.org/10.1007/s10579-021-09574-0
20. Fairclough, N.: Critical Discourse Analysis: The Critical Study of Language. Longman Applied Linguistics. Taylor & Francis (2013). https://doi.org/10.4324/9781315834368
21. Fairclough, N.: Language and Power. Language In Social Life. Taylor & Francis (2013). https://doi.org/10.4324/9781315838250
22. Fišer, D., Lenardič, J.: CLARIN resources for parliamentary discourse research. In: Fišer, D., Eskevich, M., de Jong, F. (eds.) Proceedings of LREC. ELRA (2018)
23. Forbes, M., Hwang, J.D., Shwartz, V., Sap, M., Choi, Y.: Social chemistry 101: learning to reason about social and moral norms. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of EMNLP, pp. 653–670. ACL (2020). https://doi.org/10.18653/v1/2020.emnlp-main.48
24. Fröbe, M., et al.: Continuous integration for reproducible shared tasks with TIRA.io. In: Kamps, J., et al. (eds.) ECIR 2023. LNCS, vol. 13982, pp. 236–241. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-28241-6_20
25. García-Díaz, J.A., et al.: Overview of PoliticES 2022: Spanish author profiling for political ideology. Procesamiento del Lenguaje Natural **69**, 265–272 (2022). https://doi.org/10.26342/2022-69-23
26. Gerber, C.: Gerber at touché: ideology and power identification in parliamentary debates 2024. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
27. Gerrish, S., Blei, D.M.: Predicting legislative roll calls from text. In: Getoor, L., Scheffer, T. (eds.) Proceedings of ICML, pp. 489–496. Omnipress (2011)
28. Glavaš, G., Nanni, F., Ponzetto, S.P.: Computational analysis of political texts: bridging research efforts across communities. In: 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, pp. 18–23. ACL (2019). https://doi.org/10.18653/v1/P19-4004
29. Goyal, N., Du, J., Ott, M., Anantharaman, G., Conneau, A.: Larger-scale transformers for multilingual masked language modeling. In: Rogers, A., et al. (eds.) Proceedings of RepL4NLP@ACL-IJCNLP, pp. 29–33. ACL (2021). https://doi.org/10.18653/V1/2021.REPL4NLP-1.4
30. Grancea, I.: Types of visual arguments. Argumentum. J. Seminar Discursive Logic Argumentation Theory Rhetoric **15**(2), 16–34 (2017)
31. Hariharakrishnan, J., Mirunalini, P.: Pixel phantoms at touché: ideology and power identification in parliamentary debates using linear SVC. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference

and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
32. He, P., Liu, X., Gao, J., Chen, W.: DeBERTa: decoding-enhanced BERT with disentangled attention. In: Proceedings of ICLR (2021). https://openreview.net/forum?id=XPZIaotutsD
33. Janusko, T., Kämpf, A., Keiling, D., Knick, J., Thiele, D.S.M.: Htw-dil at touché: multimodal dense information retrieval for arguments. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
34. Kiesel, J., et al.: SCaLAR NITK at touché: human value detection. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
35. Karpukhin, V., et al.: Dense passage retrieval for open-domain question answering. In: Proceedings of EMNLP, pp. 6769–6781. ACL (2020). https://doi.org/10.18653/v1/2020.emnlp-main.550
36. Khurshid, A., Das, D., Khaskel, R., Datta, S.: JU_NLP_DID at touché. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
37. Kiesel, J., Alshomary, M., Handke, N., Cai, X., Wachsmuth, H., Stein, B.: Identifying the human values behind arguments. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of ACL, pp. 4459–4471. ACL (2022). https://doi.org/10.18653/v1/2022.acl-long.306
38. Kiesel, J., et al.: SemEval-2023 task 4: ValueEval: identification of human values behind arguments. In: Kumar, R., Ojha, A.K., Doğruöz, A.S., Martino, G.D.S., Madabushi, H.T. (eds.) Proceedings of SemEval, pp. 2287–2303. ACL (2023). https://doi.org/10.18653/v1/2023.semeval-1.313
39. Kjeldsen, J.E.: Virtues of visual argumentation: how pictures make the importance and strength of an argument salient (2013)
40. Kurtoğlu Eskişar, G.M., Çöltekin, Ç.: Emotions running high? A synopsis of the state of Turkish politics through the ParlaMint corpus. In: Fišer, D., Eskevich, M., Lenardič, J., de Jong, F. (eds.) Proceedings of ParlaCLARIN, pp. 61–70. ELRA (2022). https://aclanthology.org/2022.parlaclarin-1.10
41. Legkas, S., Christodoulou, C., Zidianakis, M., Koutrintzes, D., Petasis, G., Dagioglou, M.: Hierocles of alexandria at touché: multi-task & multi-head custom architecture with transformer-based models for human value detection. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
42. Lenardič, J., Fišer, D.: CLARIN Resource Families: Parliamentary Corpora (2023). https://www.clarin.eu/resource-families/parliamentary-corpora. Accessed 09 July 2024
43. Li, Y., Bubeck, S., Eldan, R., Giorno, A.D., Gunasekar, S., Lee, Y.T.: Textbooks are All You Need II: phi-1.5 technical report (2023). https://arxiv.org/abs/2309.05463
44. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023)
45. Liu, R., Jia, C., Zhang, G., Zhuang, Z., Liu, T.X., Vosoughi, S.: Second thoughts are best: learning to re-align with human values from text edits. In: Advances in Neural Information Processing Systems, vol. 35, pp. 181–196 (2022)

46. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. CoRR (2019). http://arxiv.org/abs/1907.11692
47. Mesnard, T., et al.: Gemma: open models based on Gemini research and technology (2024). https://doi.org/10.48550/arXiv.2403.08295
48. Mirunalini, P., Koushik, A., Seshan, D.: Trojan horses at touché: logistic regression for classification of political debates. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
49. Mochtak, M., Rupnik, P., Ljubešić, N.: The ParlaSent multilingual training dataset for sentiment identification in parliamentary proceedings. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of LREC, pp. 16024–16036. ELRA and ICCL (2024). https://aclanthology.org/2024.lrec-main.1393
50. Morren, M., Mishra, R.: Eric from at touché: prompts vs finetuning. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
51. Navarretta, C., Haltrup Hansen, D.: Government and opposition in Danish parliamentary debates. In: Fiser, D., Eskevich, M., Bordon, D. (eds.) Proceedings of ParlaCLARIN, pp. 154–162. ELRA and ICCL (2024). https://aclanthology.org/2024.parlaclarin-1.23
52. Nguyen, M.V., Lai, V.D., Veyseh, A.P.B., Nguyen, T.H.: Trankit: a light-weight transformer-based toolkit for multilingual natural language processing. In: Gkatzia, D., Seddah, D. (eds.) Proceedings of EACL, pp. 80–90. ACL (2021). https://doi.org/10.18653/v1/2021.eacl-demos.10
53. Ostrower, B., Aphiwetsa, P.: Ds@gt at touché: image search and ranking via clip and image generation. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
54. Palmqvist, O., Jiremalm, J., Picazo-Sanchez, P.: Policy parsing panthers at touché: ideology and power identification in parliamentary debates. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
55. Pla, F., Hurtado, L.F.: Political tendency identification in Twitter using sentiment analysis techniques. In: Tsujii, J., Hajic, J. (eds.) Proceedings of Coling, pp. 183–192. Dublin City University and ACL (2014). urlhttps://aclanthology.org/C14-1019
56. Preoţiuc-Pietro, D., Liu, Y., Hopkins, D., Ungar, L.: Beyond binary labels: political ideology prediction of Twitter users. In: Barzilay, R., Kan, M.Y. (eds.) Proceedings of ACL, pp. 729–740. ACL (2017). https://doi.org/10.18653/v1/P17-1068
57. Qiu, L., et al.: ValueNet: a new dataset for human value driven dialogue system. In: Proceedings of AAAI, pp. 11183–11191. AAAI Press (2022). https://doi.org/10.1609/aaai.v36i10.21368
58. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of ICML, vol. 139, pp. 8748–8763. PMLR (2021). https://proceedings.mlr.press/v139/radford21a.html
59. Rapp, C.: Aristotle's rhetoric. In: Zalta, E.N., Nodelman, U. (eds.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University (2023)

60. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: Proceedings of EMNLP, pp. 3982–3992. ACL (2019). https://doi.org/10.18653/v1/D19-1410
61. Roque, G.: Visual argumentation: a further reappraisal. In: van Eemeren, F.H., Garssen, B. (eds.) Topical Themes in Argumentation Theory, vol. 22, pp. 273–288. Springer, Cham (2012). https://doi.org/10.1007/978-94-007-4041-9_18
62. Russo, D., et al.: PoliticIT at EVALITA 2023: overview of the political ideology detection in italian texts task. In: Proceedings of EVALITA. CEUR Workshop Proceedings, vol. 3473. CEUR-WS.org (2023). https://ceur-ws.org/Vol-3473/paper7.pdf
63. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter (2020)
64. Scharfbillig, M., Ponizovskiy, V., Pasztor, Z., Keimer, J., Tirone, G.: Monitoring social values in online media articles on child vaccinations. Technical report, European Commission's Joint Research Centre, Luxembourg (2022). https://doi.org/10.2760/86884
65. Scharfbillig, M., et al.: Values and identities - a policymaker's guide. Technical report, European Commission's Joint Research Centre, Luxembourg (2021). https://doi.org/10.2760/349527
66. Schwartz, S.H.: Are there universal aspects in the structure and contents of human values? J. Soc. Issues 19–45 (1994). https://doi.org/10.1111/j.1540-4560.1994.tb01196.x
67. Schwartz, S.H., et al.: Refining the theory of basic individual values. J. Pers. Soc. Psychol. (2012). https://doi.org/10.1037/a0029393
68. Sevitha, S., Patel, M., Shevgoor, S.: Team hale lab at touché 2024: ideology and power identification in parliamentary debates. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
69. Shwetha, S., Kamath, S., Balaji, S., Narayanan, S.: Vayam solve Kurmaha at touché: power identification in parliamentry speeches using TFIDF vectorizer and SVM classifier. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
70. Stefanovitch, N., Piskorski, J.: Holistic inter-annotator agreement and corpus coherence estimation in a large-scale multilingual annotation campaign. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of EMNLP, pp. 71–86. ACL (2023). https://doi.org/10.18653/v1/2023.emnlp-main.6
71. Tarkka, O., et al.: Automated emotion annotation of Finnish parliamentary speeches using GPT-4. In: Fiser, D., Eskevich, M., Bordon, D. (eds.) Proceedings of ParlaCLARIN, pp. 70–76. ELRA and ICCL (2024). https://aclanthology.org/2024.parlaclarin-1.11
72. Touvron, H., et al.: LLaMA: Open and Efficient Foundation Language Models (2023). https://doi.org/10.48550/arxiv.2302.13971
73. Kiesel, J., et al.: Ssnites at touché: ideology and power identification in parliamentary debates using BERT model. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
74. Vegetti, F., Širinić, D.: Left-right categorization and perceptions of party ideologies. Polit. Behav. **41**(1), 257–280 (2019)

75. Wachsmuth, H., et al.: Computational argumentation quality assessment in natural language. In: Proceedings of EACL, pp. 176–187 (2017). https://aclanthology.org/E17-1017
76. Yeste, V., Ardanuy, M.C., Rosso, P.: Philo of Alexandria at touché: a cascade model approach to human value detection. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
77. Yunis, H.: Arthur schopenhauer at touché 2024: multi-lingual text classification using ensembles of large language models. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
78. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of ICCV, pp. 11941–11952. IEEE Computer Society (2023). https://doi.org/10.1109/iccv51070.2023.01100
79. Zheng, L., Guha, N., Anderson, B.R., Henderson, P., Ho, D.E.: When does pretraining help?: assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings. In: Proceedings of ICAIL, pp. 159–168. ACM (2021). https://doi.org/10.1145/3462757.3466088

Author Index

L. Goeuriot et al. (Eds.): CLEF 2024, LNCS 14959, pp. 333–336, 2024.
https://doi.org/10.1007/978-3-031-71908-0