Overview of the CLEF-2024 CheckThat! Lab Task 3 on Persuasion Techniques

Jakub Piskorski[1], Nicolas Stefanovitch[2], Firoj Alam[3], Ricardo Campos[4,5], Dimitar Dimitrov[6], Alípio Jorge[7,4], Senja Pollak[8], Nikolay Ribin[6], Zoran Fijavž[9,10], Maram Hasanain[3], Purificação Silvano[7,12], Elisa Sartori[11], Nuno Guimarães[7,4], Ana Zwitter Vitez[8,13], Ana Filipa Pacheco[7], Ivan Koychev[6], Nana Yu[7], Preslav Nakov[14] and Giovanni Da San Martino[11]

[1]*Polish Academy of Sciences, Warsaw, Poland*
[2]*European Commission Joint Research Centre, Ispra, Italy*
[3]*Qatar Computing Research Institute, HBKU, Qatar*
[4]*INESC TEC, Portugal*
[5]*University of Beira Interior, Portugal*
[6]*Sofia University, Bulgaria*
[7]*University of Porto, Portugal*
[8]*Jožef Stefan Institute, Ljubljana, Slovenia*
[9]*Jožef Stefan International Postgraduate School, Ljubljana, Slovenia*
[10]*Peace Institute, Ljubljana, Slovenia*
[11]*University of Padova, Italy*
[12]*CLUP, Portugal*
[13]*University of Ljubljana, Slovenia*
[14]*Mohamed bin Zayed University of Artificial Intelligence, UAE*

Abstract

We present an overview of `CheckThat!` Lab's 2024 Task 3, which focuses on detecting 23 persuasion techniques at the text-span level in online media. The task covers five languages, namely, Arabic, Bulgarian, English, Portuguese, and Slovene, and highly-debated topics in the media, e.g., the Isreali–Palestian conflict, the Russia–Ukraine war, climate change, COVID-19, abortion, etc. A total of 23 teams registered for the task, and two of them submitted system responses which were compared against a baseline and a task organizers' system, which used a state-of-the-art transformer-based architecture. We provide a description of the dataset and the overall task setup, including the evaluation methodology, and an overview of the participating systems. The datasets accompanied with the evaluation scripts are released to the research community, which we believe will foster research on persuasion technique detection and analysis of online media content in various fields and contexts.

Keywords

Persuasion technique, media analysis, multilinguality.

1. Introduction

Fact-checking, verification and analysis of multimodal and multigenre content are of paramount importance for the reliability of information shared through various communication channels such as news, political debates, and social media. It can help prevent the spread of misinformation and promote informed decision-making. By verifying the claims in such content, individuals and organisations can make well-informed judgments and contribute to a more trustworthy online discourse.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
jpiskorski@gmail.com (J. Piskorski); Nicolas.Stefanovitch@ec.europa.eu (N. Stefanovitch); fialam@hbku.edu.qa (F. Alam); ricardo.campos@ubi.pt (R. Campos); ilijanovd@fmi.uni-sofia.bg (D. Dimitrov); amjorge@fc.up.pt (A. Jorge); senja.pollak@ijs.si (S. Pollak); n.m.ribin@gmail.com (N. Ribin); zoran.fijavz@mirovni-institut.si (Z. Fijavž); mhasanain@hbku.edu.qa (M. Hasanain); msilvano@letras.up.pt (P. Silvano); elisa.sartori.2@unipd.it (E. Sartori); nuno.r.guimaraes@inesctec.pt (N. Guimarães); Ana.ZwitterVitez@ff.uni-lj.si (A. Zwitter Vitez); anafilipasrpacheco@gmail.com (A. F. Pacheco); koychev@fmi.uni-sofia.bg (I. Koychev); robertananayu@hotmail.com (N. Yu); preslav.nakov@mbzuai.ac.ae (P. Nakov); giovanni.dasanmartino@unipd.it (G. Da San Martino)

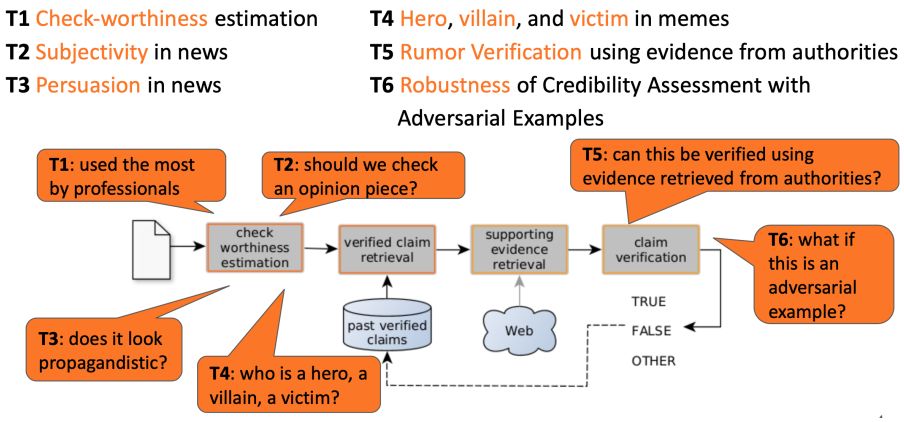

Figure 1: The CheckThat! lab verification pipeline. The 2024 edition of the lab covers six tasks: (*T1*) check-worthiness estimation, (*T2*) subjectivity, (*T3*) persuasion technique detection (this paper), (*T4*) detecting hero, villain, and victim from memes, (*T5*) rumour verification using evidence from authorities, and (*T6*) robustness of credibility assessment with adversarial examples.

This paper offers an overview of the shared task on detecting teh use of persuasion techniques in multilingual news which was organized as part of CheckThat! 2024 lab. The CheckThat! 2024 lab was held in the framework of CLEF 2024 [1].[1] Figure 1 shows the full CheckThat! identification and verification pipeline, highlighting the six tasks targeted in this seventh edition of the lab: Task 1 on check-worthiness estimation, Task 2 on subjectivity, Task 3 on persuasion technique detection (this paper), Task 4 on detecting hero, villain, and victim in memes, Task 5 on rumor verification using evidence from authorities, and Task 6 on robustness of credibility assessment with adversarial examples.

Task 3 focuses on the detection of 23 persuasion techniques at text-span level in online media. The task covers 5 languages, namely, Arabic, Bulgarian, English, Portuguese, and Slovene and highly-debated topics in the media. A total of 23 teams registered for the task, and two of them submitted systems, which were compared against a baseline and a task organizers' system, which uses a state-of-the-art transformed-based architecture. The participating systems also used state-of-the-art transformer-based architectures and data augmentation.

The remainder of this paper is organized as follows: Section 2 briefly presents the task. Section 3 describes the datasets and the evaluation methodology. Section 5 gives an overview of the system submissions, the organizers' system, and the evaluation results. Section 6 presents related work, whereas Section 7 offers some final conclusions.

2. Task

The goal of this task is to recognize and classify persuasion techniques in multilingual news at the text span level. In particular, we exploit the two-tier persuasion technique taxonomy introduced in *SemEval 2023 Shared Task 3 on Detecting the Genre, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup* [2]. At the top level, there are 6 coarse-grained types of persuasion techniques: *Attack on reputation*, *Justification*, *Simplification*, *Distraction*, *Call*, and *Manipulative wording*. These six main types are further subdivided into 23 fine-grained techniques. Figure 2 presents the entire taxonomy. Figure 3 provides one example of persuasion technique per main category. Full definitions and further examples of persuasion techniques are given in Piskorski et al. [3] and Piskorski et al. [4].

As training and development data, we used the existing corpus from the aforementioned SemEval 2023 task [2], which covers nine languages: English, German, Georgian, Greek, French, Italian, Polish, Russian, and Spanish. For test data, we developed an entirely new dataset that covers five languages: Arabic, Bulgarian, English, Portuguese, and Slovene. English is the only language for which both training/development and test data exist.

[1] https://checkthat.gitlab.io/

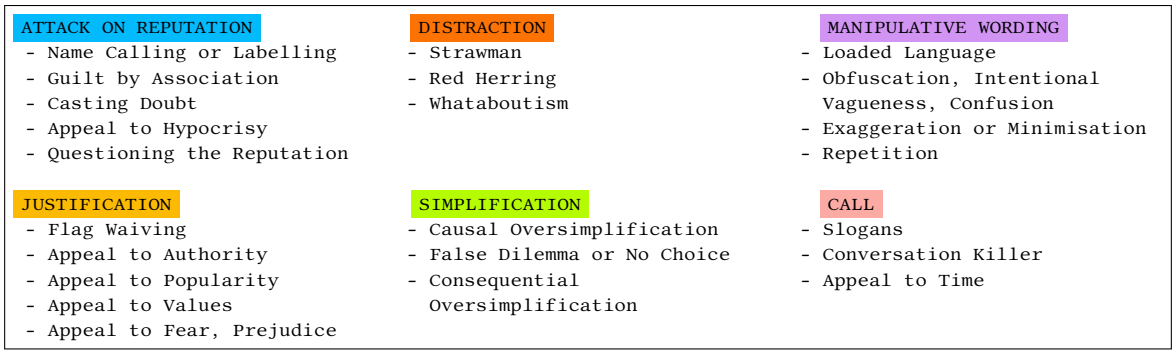

Figure 2: Two-tier persuasion technique taxonomy.

Name Calling or Labelling: *'Fascist' **Anti-Vax** Riot Sparks COVID Outbreak in Australia.*

Appeal to Authority: ***Since the Pope said that this aspect of the doctrine is true we should add it to the creed.***

Strawman: *Referring to your claim that providing medicare for all citizens would be costly and a danger to the free market, I infer **that you don't care if people die from not having healthcare, so we are not going to support your endeavour**.*

Consequential Oversimplification: ***If we begin to restrict freedom of speech, this will encourage the government to infringe upon other fundamental rights, and eventually this will result in a totalitarian state where citizens have little to no control of their lives and decisions they make***

Slogans: ***"Immigrants welcome, racist not!***

Exaggeration or Minimisation: *From the seminaries, to the clergy, to the bishops, to the cardinals, **homosexuals are present at all levels, by the thousand***

Figure 3: Examples of text snippets with persuasion techniques. The text fragments highlighted in bold are the actual text spans annotated.

The test dataset covers highly-debated topics in the media, e.g., the Isreali-Palestian conflict, the Russia–Ukraine war, climate change, COVID-19, abortion, etc. Except for the Israeli-Palestinian conflict, the same topics are also covered in the training/development dataset. For Arabic, the dataset covers forteen broad topics such as news, politics, health, social, sports, arts and culture, religion, science and technology, human rights, and lifestyle. Among them, *news* and *politics* cover more than 50% of the paragraphs. More detail about the topic distribution can be found in [5].

The main difference between the Task 3 presented in this paper and the former competition on persuasion technique detection organized at SemEval 2023 [2] is that the latter focused on the detection of persuasion techniques at the paragraph level, while the current task aims at developing models to detect and to classify persuasion techniques at the span level, which constitutes an additional challenge.

3. Datasets

3.1. Annotation Process

Each language was annotated by a team of annotators fluent in the language and used to perform such annotations; the language leaders met regularly in order to discuss difficult cases with more experienced annotators having already taken part in previous annotations campaign using the same taxonomy. For all languages but Arabic, each document was annotated by two annotators, and one curator reconciled the annotations. For the Arabic test dataset, each paragraph was annotated by three annotators, and two curators consolidated the annotations.

We followed the approach laid in [4] to train annotators, with the exception of Arabic: they were first given the comprehensive annotation guidelines, were further trained using two sets of flashcards of increasing complexity, and lastly had to annotate and to discuss with expert annotators five test documents whose ground-truth annotations were known.

3.2. Quality and Coherence Assurance

The overal Inter-Annotator Agreement (IAA) as measured by the Krippendorff's α is of 0.404, which is lower than the recommended value of 0.667. In Table 2, we also reported the α for each language independently. One has to take into account that this measures coherence before curation, and that significant steps have been taken in order to improve the quality of the curated data, as described below.

We used the approach of [6], which facilitates the comparison of annotations across documents and across languages. This allowed us to cluster the annotations based on their semantic similarity, which was used to flag outliers for review, and allowed us to spot cross-lingual disagreements. Such disagreements were either due to individual annotator differences or to a more fundamental different understanding of techniques' definitions across language-specific annotation teams. Such differences could concern either nuances of the meaning of specific labels or the length of the span to be selected.

We further used the following additional measures to improve the quality of the dataset: (a) we compared the distribution of labels to spot obvious cross-lingual inconsistencies, (b) we alphabetically sorted texts in order to make it easy to spot similar texts with different labels, and (c) finally the most experienced annotators did random checks.

All these measures contributed to the increase of the coherence of the dataset, which could be measured for all languages except for Arabic using the o value as defined in [6] and using the same settings: when ignoring the *Loaded Language* and *Name-Calling Labelling* classes, the value goes from 0.279 to 0.284, and when considering them it increases from 0.608 to 0.611; this increase is mostly driven by improvement of the inter-language coherence.

3.3. Statistics about the Datasets

3.3.1. Training and Development Data

The overall statistics about the training and the development datasets are provided in Table 1. For more detailed characteristics of these datasets, please refer to [2] and Piskorski et al. [3].

Table 1
Training and *development* dataset statistics.

	Training		**Development**	
language	#documents	#spans	#documents	#spans
English	536	9,002	54	1,775
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

3.3.2. Test Dataset

As part of Task 3, we created new labeled datasets for Arabic, Bulgarian, English, Portuguese, and Slovene. With the exception of the latter, this shared task is the first application of the framework for annotating persuasion techniques for the mentioned languages. News selection was delegated to the teams responsible for their respective languages, but they were expected to include a variety of topics, news genres, and political stances, in addition to selecting texts where a high prevalence of persuasion techniques was to be expected. To allow for comparability with previous datasets, the topics of the Russia–Ukraine war, climate change, COVID-19, and abortion were covered in all test datasets except for Arabic. In addition, a new topic, the Israeli–Palestinian conflict, was added.

The number of included news articles and the topic distributions for the languages are presented in Tables 2 and 3, respectively. Overall, the most commonly annotated persuasion technique was *Loaded language*, followed by *Name-calling*, *Casting doubt* and *Questioning the Reputation*, although the specific distribution varies across the datasets. The share of annotated persuasion technique classes across the test datasets is presented in Table 4. The distribution of the frequency of the persuasion techniques in the test dataset is to some degree similar and comparable to the datasets used in the SemEval 2023 Task on persuasion techniques [2], i.e., *Loaded language* and *Name-calling* are the two most prevalent fine-grained techniques, whereas *Manipulative Wording* and *Attack on Reputation* are the two coarse-grained persuasion technique categories with highest share in both datasets.

Table 2
Test dataset statistics. Note that for Arabic, the number of articles does not directly reflect the number of paragraphs, as we only annotated soem selected paragraphs from them.

	Test			
language	#documents	#paragraphs	#spans	α
Arabic	1,527	1,642	2,197	-
Bulgarian	100	916	1,732	0.197
English	98	2,174	2,599	0.168
Slovenian	100	1,478	4,591	0.470
Portuguese	104	1,501	1,727	0.587

Table 3
Test dataset topic distributions.

	Topics (%)						
language	Israel-Palestine	Ukraine-Russia	Covid-19	Migration	Climate	Abortion	Elections
Arabic	-	-	-	-	-	-	-
Bulgarian	19.0	13.0	15.0	23.0	15.0	15.0	-
English	25.5	13.3	15.3	15.3	15.3	15.3	-
Slovenian	20.0	20.0	16.0	15.0	15.0	14.0	-
Portuguese	24.0	18.3	10.6	10.6	12.5	11.5	12.5

The Portuguese dataset was developed by manually selecting 104 articles from 28 European Portuguese news media based on the main topics of the task. Moreover, a combination of news and opinion pieces was selected to ensure a variety of annotation types. The Portuguese dataset consists of articles from 27 different news sources, where 25 articles are related to the Israeli–Palestian conflict, 19 to the Russia–Ukraine war, 11 to COVID-19, 11 to migration, 13 to climate change, 12 to abortion and 13 to elections. The election topic was included due to the large number of news and opinion articles released on this topic during the extraction process. In addition, these articles are rich in persuasion techniques, making them a good fit for the current task.

The Slovenian dataset included manually selected 100 news articles from 11 news channels and two blogs, with the latter being used to preserve a balance across political leanings and topics. Hard news constituted 20% of the included text with the rest consisting of opinion articles. Topically, 20 of the annotated articles were related to the Israeli-Palestinian conflict, 20 to the Russia–Ukraine war, 16 to COVID-19, 15 to climate change, 15 to migration, and 14 to discussions on gender-related topics. The latter was done as the right to abortion is contitutionally protected in Slovenia and rarely contested directly.

The English dataset has a total of 98 articles from 80 unique news sources. Of the 98 articles, 25 articles concern the Israeli-Palestinian conflict, 15 the topic of climate change, 15 the abortion discussion, 15 COVID-19, 15 migration, and 13 the Russia–Ukraine war. The articles were collected manually using both news and opinion articles from media outlets present in Media Bias/Fact Check[2].

[2]https://mediabiasfactcheck.com

Table 4
Distribution of persuasion technique labels in test dataset by language (percent). Color intensity represents the relative frequency of labels for each language.

		Persuasion Techniques Distribution by Language (%)				
		Arabic	**Bulgarian**	**English**	**Slovenian**	**Portuguese**
Attack on Reputation	Name-Calling Labeling	17,60	11,10	15,90	14,70	12,40
	Guilt by Association	0,20	2,90	1,70	1,40	0,60
	Doubt	1,40	10,20	8,60	12,70	6,30
	Appeal to Hypocrisy	0,30	0,90	1,50	0,30	1,00
	Questioning the Reputation	3,70	5,50	15,00	14,20	19,10
Justification	Flag Waving	1,00	1,80	2,30	0,50	0,90
	Appeal to Authority	0,50	1,40	5,10	3,70	2,60
	Appeal to Popularity	0,10	0,90	1,00	0,90	0,60
	Appeal to Values	0,40	4,30	4,20	3,70	8,00
	Appeal to Fear-Prejudice	0,50	8,40	7,00	5,00	7,20
Distraction	Strawman	0,10	2,00	0,70	1,90	0,30
	Red Herring	0,20	0,90	0,50	0,60	0,50
	Whataboutism	0,00	1,10	1,70	0,10	0,20
Simplification	Causal Oversimplification	1,20	1,80	3,10	2,80	2,10
	False Dilemma-No Choice	0,10	2,70	4,80	3,40	3,00
	Consequential Oversimplification	0,30	1,40	1,80	1,80	3,40
Manipulative Wording	Loaded Language	60,90	22,80	16,90	25,90	15,10
	Obfuscation-Vagueness-Confusion	2,30	0,40	0,20	0,20	0,90
	Exaggeration-Minimisation	7,30	8,10	0,90	1,50	2,70
	Repetition	0,50	6,20	2,20	1,60	6,10
Call	Slogans	0,70	2,90	2,30	1,40	1,60
	Conversation Killer	0,30	1,40	1,60	0,90	2,30
	Appeal to Time	0,40	0,80	1,00	0,80	3,10

The Arabic dataset consists of Arabic news articles from AraFacts [7] and an in-house news article collection. We split the articles into paragraphs and annotated them at the paragraph level. From the AraFacts news articles, we annotated all paragraphs, while from the in-house news article collection, we randomly selected paragraphs by stratified sampling over news media, ensuring diversity in topics and news media. The dataset covers around 14 broad topics, with news and politics being the top most frequently covered ones. More details about this dataset can be found in [5].

The Bulgarian dataset consists of 100 manually selected articles extracted from 9 different sources. There are 19 articles on the Israeli-Palestinian conflict, 18 on COVID-19, 15 on climate change, 25 on the Russia-Ukraine war, and 23 on migration.

4. Evaluation Framework

Task Organization

For the lab, we provided training and development datasets. The latter was intended to allow participants to validate their systems internally, while they could use the development set for hyper-parameter tuning and model selection. The test set was used for the final evaluation and ranking. The participants were allowed to submit multiple runs on the test set (without seeing the scores), but only the last valid run was considered as their official submission.

Evaluation Measure

The task is defined as a multi-label multi-class sequence tagging problem, as such traditional evaluation metrics tend to be too strict when scoring since they are based on exact matching. We analyzed several annotated articles and how the spans varied between different annotators and consolidators and discovered that most of the time there was agreement on the technique, but the spans differed slightly.

It is also important to emphasize that from the end-user perspective (e.g., analysts carrying out comparative media analysis) partial matches with significant overlap could be considered as equally good as exact matches. To address the limitations of exact matching scorers, we propose an adjustment to the traditional F_1-score to take into account partial span matching. Let

- P = {p_1, ..., p_n} be the set of predictions for one article, $p \in P$ is a generic prediction which is represented as an ordered triple $\langle span_{\text{start}}, span_{\text{end}}, label \rangle$
- G = {g_1, ..., g_m} be the set of gold labels for one article, $g \in G$ is a generic gold label which is represented as an ordered triple $\langle span_{\text{start}}, span_{\text{end}}, label \rangle$
- $L : (p, g) \longrightarrow \{0, 1\}$ is a function that measures the similarity of the labels of p and g

$$L(p, g) = \begin{cases} 1, & \text{if the labels of p and g are identical} \\ 0, & \text{otherwise} \end{cases}$$

- $I : (p, g) \longrightarrow [0, 1]$ is a function that measures the overlap rate of the spans of p and g

$$I(p, g) = \begin{cases} 1, & \text{if } \frac{|p \cap g|}{|g|} \geqslant 0.5 \text{ and } |p| \leqslant 2 \cdot |g| \\ \frac{|p \cap g|}{|g|} \in (0, 1), & \text{if } \frac{|p \cap g|}{|g|} \in (0, 0.5) \text{ and } |p| \leqslant 2 \cdot |g| \\ \frac{|p \cap g|}{|p|} \in (0, 1), & \text{if } \frac{|p \cap g|}{|g|} \in (0, 1] \text{ and } |p| > 2 \cdot |g| \text{ and } |p| \leqslant 4 \cdot |g| \\ 0, & \text{otherwise} \end{cases}$$

- $S : (p, g) \longrightarrow [0, 1]$ is a similarity function of two spans p and g. It is calculated as follows:

$$S(p, g) = L(p, g) \cdot I(p, g)$$

We map each possible case into True Positive (T_p), False Positive (F_p), and False Negative (F_n) values, and then compute the standard F_1-score. From the obtained values of T_p, F_p, F_n, we compute the F_1-score of the example. Additionally, F_1-score is computed for each persuasion technique. For all datasets, the results are micro- and macro-averaged.

The pseudocode of the algorithm for computing *True Positives, False Positives, and False Negatives* is given in Algorithm 1.

Baseline System

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 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. We used XLM-RoBERTa-base [8] in a zero-shot setting.

5. Results and Overview of the Systems

The task is 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.

Algorithm 1 Pseudocode for the evaluation measure of the task.

Let M be an empty list of pairs, where each element $\langle p, g \rangle$ is a pair of prediction and a gold label
while $P \neq \varnothing$ **and** $G \neq \varnothing$ **do**
    find $\langle p*, g* \rangle$ which maximises $S(p, g)$
    $M \leftarrow M \cup \langle p*, g* \rangle$
    $P \leftarrow P \backslash \{p*\}$
    $G \leftarrow G \backslash \{g*\}$
end while
$F_{\mathrm{n}} \leftarrow |G|$ ▷ gold labels left with no match are false negatives
$F_{\mathrm{p}} \leftarrow |P|$ ▷ predictions left with no match (or already matched with the same or better similarity value) are false positives
$T_{\mathrm{p}} \leftarrow 0$
for each $\langle p, g \rangle \in M$ **do**
    $T_{\mathrm{p}} \leftarrow T_{\mathrm{p}} + S(p, g)$
    $F_{\mathrm{p}} \leftarrow F_{\mathrm{p}} + (1 - S(p, g))$ ▷ partial given credit affects the score depending on the prediction mistake
end for

Table 5
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 6
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				

5.1. Participating Systems

In Table 5, we provide an overview of 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 6, we report the results. Team **UniBO** participated in all languages but ranked first only for English. Team **Mela** participated only in Arabic and was the top-ranked system, showing a significant improvement compared to other teams and the baseline.

Team **UniBO** [9] 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** [10] proposed a multilingual BERT-based system that incorporates both English and Arabic knowledge during its pre-training stage. With this system, they achieved first place on the Arabic leaderboard in the shared task.

5.2. Organizer System

For the sake of comparison to state-of-the-art solutions, we as organizers developed (after the competition) 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 [11] trained on the SemEval-2023 corpus [3], capable of processing arbitrarily long text using sliding window chunking with 50% overlap. This classifier achieves state-of-the-art results on the SemEval 2023 Task 3 test dataset [2] for all six languages (oscillates around 1-3 rank across languages), both in terms of micro and macro F_1 scores. Further detail about this classifier can be found in [12].

6. Related Work

Early work on persuasion techniques focused on one or a few specific ones: Habernal et al. [13, 14] developed a corpus with 1.3k arguments annotated with five fallacies. Da San Martino et al. [15], created a corpus of news articles annotated with 18 techniques, considering separately the task of technique spans detection and classification. They further tackled a sentence-level propaganda detection task, and proposed a multi-granular gated deep neural network. The model was afterwards implemented in the Prta system [16], while the data was expanded to include nine languages [3]. Several models were proposed to address the limitations of transformers [17], or looking into interpretable propaganda detection [18]. Persuasion techniques are used also in memes, requiring complex multimodal models [19]. The survey on computational propaganda detection [20] highlights the need for models combining NLP and Network Analysis, which has been further analysed in Hristakieva et al. [21]. Other works analysed also the use of persuasion techniques on social media posts about COVID-19 [22, 23].

Several shared tasks on the detection of persuasion techniques have been organised through the years: the *NLP4IF-2019 task on Fine-Grained Propaganda Detection* [24], which is based on a subset of the data of this task; the *SemEval-2020 task 11 on Detection of Persuasion Techniques in News Articles* [25], which considers 14 techniques and split the task of identifying any persuasive span and, given the span, identify the technique in it; the *SemEval-2021 task 6 on Detection of Persuasion Techniques in Texts and Images* focused on 22 techniques in memes [26]; the *WANLP'2022 shared task* asked to detect the use of 20 propaganda techniques in Arabic tweets [27]; the *SemEval-2023 Task 3* includes several languages and defines the problem as a multilabel classification one at paragraph level [2]; *Task 1 at the ArAIEval shared task* targets the same task and techniques of our shared task, covering multi-genre Arabic content [28]. Further annotations in Spanish and English, although for different categories of techniques, are provided by the DIPROMATS initiative [29].

7. Conclusion and Future Work

We presented an overview of task 3 of the CLEF-2024 `CheckThat!` lab. The lab featured tasks that span the full verification pipeline: from spotting check-worthy claims to claim verification. Task 3 focused on the detection of 23 persuasion techniques at the text span level in online media, and covers five languages (Arabic, Bulgarian, English, Portuguese, and Slovene) and highly-debated topics in the media.

The task is a natural follow-up of the former competition on persuasion technique detection organized at SemEval-2023 [2], which focused on the detection of persuasion techniques at the paragraph level, while the task described in this paper aims at developing models to detect and to classify persuasion techniques at the span level, constituting an additional complexity. The participating systems used state-of-the-art transformer-based architectures and deployed data augmentation. The obtained results compared vis-a-vis the results reported in [2] confirm that the detection at the span level is a harder task, and leaves space for improvement. In future work, we plan to expand the task in a variety of ways, e.g., by enlarging the dataset, by incorporating more languages, and by considering other text genre, e.g., parliamentary debates.

Limitations While creating the test dataset used in this task, we strived to have a balanced representation of the points of view on the various topics, but this was done on a best-effort basis, and the data might not be fully representative for carrying out other type of research, e.g., comparative media analysis, etc.

Acknowledgments

We are greatly indebted to the following persons who contributed to the organization of this task: Nikolaos Nikolaidis, Ivanka Mavrodieva, Desislava Angelova, Ana Rupnik, Anja Krivec, Ita Osredkar, Katja Štefanič, Lana Valič, Maja Habjanič, Rok Drenik, Špela Rot, Veronika Razpotnik, Zala Roguljič.

The work of F. Alam, M. Hasanain and G. Da San Martino is partially supported by NPRP 14C-0916-210015 from the Qatar National Research Fund, which is a part of Qatar Research Development and Innovation Council (QRDI). The findings achieved herein are solely the responsibility of the authors.

The work of R.Campos, A. Jorge, and P. Silvano was financed by National Funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology) within the project StorySense, with reference 2022.09312.PTDC (DOI 10.54499/2022.09312.PTDC).

The work of N. Guimarães was financed by Component 5 - Capitalization and Business Innovation, integrated in the Resilience Dimension of the Recovery and Resilience Plan within the scope of the Recovery and Resilience Mechanism (MRR) of the European Union (EU), framed in the Next Generation EU, for the period 2021 - 2026, within project HfPT, with reference 41.

The work of D. Dimitrov, N. Ribin, and I. Koychev is partially financed by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project SUMMIT, No BG-RRP-2.004-0008.

The work of S. Pollak, Z. Fijavž, and A. Zwitter Vitez was supported by the Slovenian Research Agency grants via the core research programmes Knowledge Technologies (P2-0103), Equality and Human Rights in the Times of Global Governance (P5-0413) and Theoretical and Applied Linguistic Research: Contrastive, Synchronic, and Diachronic aspects (P6-0218), and the projects Computer-assisted multilingual news discourse analysis with contextual embeddings (J6-2581), Embeddings-based techniques for Media Monitoring Applications (L2-50070), Hate Speech in Contemporary Conceptualizations of Nationalism, Racism, Gender and Migration (J5-3102). .

References

[1] A. Barrón-Cedeño, F. Alam, J. M. Struß, P. Nakov, T. Chakraborty, T. Elsayed, P. Przybyła, T. Caselli, G. Da San Martino, F. Haouari, C. Li, J. Piskorski, F. Ruggeri, X. Song, R. Suwaileh, Overview of the CLEF-2024 CheckThat! Lab: Check-worthiness, subjectivity, persuasion, roles, authorities and adversarial robustness, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. Di Nunzio, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[2] J. Piskorski, N. Stefanovitch, G. Da San Martino, P. Nakov, SemEval-2023 Task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Toronto, Canada, 2023, pp. 2343–2361.
[3] J. Piskorski, N. Stefanovitch, N. Nikolaidis, G. Da San Martino, P. Nakov, Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada, 2023, pp. 3001–3022.
[4] J. Piskorski, N. Stefanovitch, V.-A. Bausier, N. Faggiani, J. Linge, S. Kharazi, N. Nikolaidis, G. Teodori, B. De Longueville, B. Doherty, J. Gonin, C. Ignat, B. Kotseva, E. Mantica, L. Marcaletti, E. Rossi, A. Spadaro, M. Verile, G. Da San Martino, F. Alam, P. Nakov, News Categorization, Framing and Persuasion Techniques: Annotation Guidelines, Technical Report, European Commission Joint Research Centre, Ispra (Italy), 2023.
[5] M. Hasanain, F. Ahmad, F. Alam, Can GPT-4 identify propaganda? annotation and detection of propaganda spans in news articles, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italia, 2024, pp. 2724–2744.
[6] N. Stefanovitch, J. Piskorski, Holistic inter-annotator agreement and corpus coherence estimation in a large-scale multilingual annotation campaign, in: Conference on Empirical Methods in Natural Language Processing, 2023.
[7] Z. Sheikh Ali, W. Mansour, T. Elsayed, A. Al-Ali, AraFacts: the first large arabic dataset of naturally occurring claims, in: Proceedings of the Sixth Arabic Natural Language Processing Workshop, 2021, pp. 231–236.
[8] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451.
[9] P. Gajo, L. Giordano, A. Barron-Cedeño, UniBO at CheckThat! 2024: Multi-lingual and Multi-label Persuasion Technique Detection in News with Data Augmentation and Sequence-Token Classifiers, in: [30], 2024.
[10] S. Nabhani, M. A. R. Riyadh, Mela at CheckThat! 2024: Transferring Persuasion Detection from English to Arabic - A Multilingual BERT Approach, in: [30], 2024.
[11] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019).
[12] N. Nikolaidis, J. Piskorski, N. Stefanovitch, Exploring the usability of persuasion techniques for downstream misinformation-related classification tasks, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italia, 2024, pp. 6992–7006.
[13] I. Habernal, R. Hannemann, C. Pollak, C. Klamm, P. Pauli, I. Gurevych, Argotario: Computational argumentation meets serious games, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP '17, Copenhagen, Denmark, 2017, pp. 7–12.
[14] I. Habernal, P. Pauli, I. Gurevych, Adapting serious game for fallacious argumentation to German: Pitfalls, insights, and best practices, in: LREC, 2018.
[15] G. Da San Martino, S. Yu, A. Barron-Cedeno, R. Petrov, P. Nakov, Fine-grained analysis of propaganda in news articles, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, EMNLP '19, Hong Kong, China, 2019, pp. 5636–5646.
[16] G. Da San Martino, S. Shaar, Y. Zhang, S. Yu, A. Barrón-Cedeno, P. Nakov, Prta: A system to support the analysis of propaganda techniques in the news, in: ACL, 2020.

[17] A. Chernyavskiy, D. Ilvovsky, P. Nakov, Transformers: "The end of history" for NLP?, in: Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML-PKDD'21, 2021.

[18] S. Yu, G. Da San Martino, M. Mohtarami, J. Glass, P. Nakov, Interpretable propaganda detection in news articles, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP '21, 2021, pp. 1597–1605.

[19] D. Dimitrov, B. B. Ali, S. Shaar, F. Alam, F. Silvestri, H. Firooz, P. Nakov, G. Da San Martino, Detecting propaganda techniques in memes, in: Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL-IJCNLP '21, 2021, pp. 6603–6617.

[20] G. Da San Martino, S. Cresci, A. Barrón-Cedeño, S. Yu, R. Di Pietro, P. Nakov, A survey on computational propaganda detection, in: C. Bessiere (Ed.), Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI-PRICAI '20, 2020, pp. 4826–4832.

[21] K. Hristakieva, S. Cresci, G. Da San Martino, M. Conti, P. Nakov, The spread of propaganda by coordinated communities on social media, in: Proceedings of the 14th ACM Web Science Conference, WebSci '22, Barcelona, Spain, 2022, pp. 191–201.

[22] P. Nakov, F. Alam, S. Shaar, G. Da San Martino, Y. Zhang, COVID-19 in Bulgarian social media: Factuality, harmfulness, propaganda, and framing, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP '21, 2021.

[23] P. Nakov, F. Alam, S. Shaar, G. Da San Martino, Y. Zhang, A second pandemic? Analysis of fake news about COVID-19 vaccines in Qatar, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP '21, 2021.

[24] G. Da San Martino, A. Barrón-Cedeño, P. Nakov, Findings of the NLP4IF-2019 shared task on fine-grained propaganda detection, in: Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, NLP4IF '19, Hong Kong, China, 2019, pp. 162–170.

[25] G. Da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, P. Nakov, SemEval-2020 task 11: Detection of propaganda techniques in news articles, in: Proceedings of the International Workshop on Semantic Evaluation, SemEval '20, Barcelona, Spain, 2020.

[26] D. Dimitrov, B. Bin Ali, S. Shaar, F. Alam, F. Silvestri, H. Firooz, P. Nakov, G. Da San Martino, Task 6 at SemEval-2021: Detection of persuasion techniques in texts and images, in: Proceedings of the 15th International Workshop on Semantic Evaluation, SemEval '21, Bangkok, Thailand, 2021.

[27] F. Alam, H. Mubarak, W. Zaghouani, G. Da San Martino, P. Nakov, Overview of the WANLP 2022 shared task on propaganda detection in Arabic, in: Proceedings of the Seventh Arabic Natural Language Processing Workshop, WANLP '22, Abu Dhabi, UAE, 2022.

[28] M. Hasanain, M. A. Hasan, F. Ahmed, R. Suwaileh, M. R. Biswas, W. Zaghouani, F. Alam, ArAIEval Shared Task: Propagandistic Techniques Detection in Unimodal and Multimodal Arabic Content, in: Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024), Bangkok, 2024.

[29] P. M. y Guillermo Marco y Julio Gonzalo y Jorge Carrillo-de-Albornoz y Iván Gonzalo-Verdugo, Overview of DIPROMATS 2023: automatic detection and characterization of propaganda techniques in messages from diplomats and authorities of world powers, Procesamiento del Lenguaje Natural 71 (2023) 397–407.

[30] G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CLEF 2024, 2024.