Overview of the 9th Social Media Mining for Health Applications (#SMM4H) Shared Tasks at ACL 2024 – Large Language Models and Generalizability for Social Media NLP

Dongfang Xu[1], Guillermo Lopez-Garcia[1], Lisa Raithel[2, 3, 4], Roland Roller[2], Philippe Thomas[2], Eiji Aramaki[5], Shoko Wakamiya[5], Shuntaro Yada[5], Pierre Zweigenbaum[6], Karen O'Connor[7], Sophia Hernandez[8], Sai Tharuni Samineni[1], Yao Ge[9], Swati Rajwal[9], Sudeshna Das[9], Abeed Sarker[9], Ari Klein[7], Ana Lucia Schmidt[10], Vishakha Sharma[11], Raul Rodriguez-Esteban[10], Juan M. Banda[12], Ivan Flores Amaro[1], Davy Weissenbacher[1], Graciela Gonzalez-Hernandez[1]

[1]Cedars-Sinai Medical Center, Los Angeles, CA, USA
[2]German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
[3] Quality & Usability Lab, Technische Universität Berlin, Berlin, Germany
[4] BIFOLD – Berlin Institute for the Foundations of Learning and Data, Berlin, Germany
[5]Nara Institute of Science and Technology, Nara, Japan
[6]Université Paris-Saclay, CNRS, LISN, Orsay, France
[7]University of Pennsylvania, Philadelphia, PA, USA
[8]University of Pittsburgh, Pittsburgh, PA, USA
[9]Emory University, Atlanta, GA, USA
[10]Roche Innovation Center, Basel, Switzerland
[11]Roche Diagnostics, Santa Clara, CA, USA
[12]Stanford Health Care, Newark, CA, USA
Correspondence: dongfang.xu@cshs.org

Abstract

For the past nine years, the Social Media Mining for Health Applications (#SMM4H) shared tasks have promoted community-driven development and evaluation of advanced natural language processing systems to detect, extract, and normalize health-related information in publicly available user-generated content. This year, #SMM4H included seven shared tasks in English, Japanese, German, French, and Spanish from Twitter, Reddit, and health forums. A total of 84 teams from 22 countries registered for #SMM4H, and 45 teams participated in at least one task. This represents a growth of 180% and 160% in registration and participation, respectively, compared to the last iteration. This paper provides an overview of the tasks and participating systems. The data sets remain available upon request, and new systems can be evaluated through the post-evaluation phase on CodaLab.

1 Introduction

The number of social media (SM) users continues to grow worldwide: 86% of US adults and 79% in Europe use SM (Center; Elliott and Sverdlov, 2012). Advances in automated data processing, machine learning and natural language processing (NLP) allow us to incorporate this massive real-time data source from around the world for biomedical and public health applications, providing researchers a venue to address the many methodological challenges unique to this media. The Social Media Mining for Health Applications (#SMM4H) Workshop, in its 9th annual iteration, brings together researchers interested in developing and sharing NLP methods that enable the systematic use of SM data for health research. The tasks of this year use data from various platforms (X, Reddit, and patient forums such as Lifeline[1] or YJQA[2]) and languages (English, Spanish, French, German, and Japanese), with a special focus on Large Language Models (LLMs) for Social Media NLP. Seven tasks organized by experienced research teams from around the world were selected for 2024. We prioritized tasks evaluating the generalizability of the proposed approaches by explicitly creating test sets with out-of-distribution data such as unseen concepts, multi-lingual and multi-source texts. These tasks are: extraction and normalization of adverse drug events in English tweets (Task 1), cross-lingual few-shot relation extraction for pharmacovigilance in French, German,

[1] https://fragen.lifeline.de/forum/
[2] https://chiebukuro.yahoo.co.jp/

and Japanese (Task 2), multi-class classification of effects of outdoor spaces on social anxiety symptoms in Reddit (Task 3), extraction of the clinical and social impacts of non-medical substance use from Reddit (Task 4), binary classification of English tweets reporting children's medical disorders (Task 5), self-reported exact age classification with cross-platform evaluation in English (Task 6), and identification of whether an LLM or a human domain expert annotated data in the context of health-related applications (Task 7).

Teams could register for a single task or multiple tasks. Teams were provided with gold-standard annotated training and validation sets to develop their systems and, subsequently, an unlabeled test set for the final evaluation. After receiving the test set, all teams were given 5 days to submit the predictions of their systems to CodaLab —a platform that facilitates data science competitions— for automatic evaluation, promoting a systematic performance comparison. Among the 84 teams that registered, 45 teams submitted at least one set of predictions: 11 teams for Task 1, 3 teams for Task 2, 15 teams for Task 3, 3 teams for Task 4, 20 teams for Task 5, 7 teams for Task 6, and 3 teams for Task 7. Teams that submitted predictions were invited to submit a short manuscript describing their system, and 38 of the 45 teams did. Each of these 38 system descriptions was peer-reviewed by at least 2 reviewers. In this article, we present the annotated corpora, the technical summaries of all the participating systems, and the performance results, providing insights into state-of-the-art methods for mining social media data for health informatics.

2 Tasks

2.1 Task 1: Extraction and normalization of adverse drug events in English tweets

Adverse drug events (ADEs) are harmful and undesired reactions attributed to the intake of a drug or medication. Active post-market surveillance is essential, given clinical trials may not detect all potential ADEs, particularly for vulnerable populations. Social media can complement traditional reporting systems, such as the FDA's Adverse Event Reporting System (FAERS) for pharmacovigilance (Leaman et al., 2010; Tricco et al., 2018). Task 1 involved automatically extracting ADE text spans in tweets and normalizing them to their standard preferred term IDs (ptIDs) in the Medical Dictionary for Regulatory Activities (MedDRA).

Dataset The dataset for Task 1 contains a total of 18,185 tweets with 1,650 adverse drug events (ADEs) labeled in the training set, 965 tweets with 85 ADEs in the development set, and 11,799 tweets with 1,232 ADEs in the test set. The training, development, and test splits include 1,239, 65, and 915 tweets reporting at least one ADE. Notably, 5.8% of the ADEs in the development set are unseen (i.e., they do not appear in the training set), and 22.0% of the ADEs in the test set are unseen (i.e., they do not appear in either the training or development sets). This was done explicitly to test the systems' generalizability (understood as their capacity to detect unseen mentions).

Evaluation We used three different evaluation metrics: the ADE normalization score for all ptIDs, the ADE normalization score for unseen ptIDs, and ADE extraction scores. The first two metrics are the same as those used in the shared task in SMM4H-2023 (Klein et al., 2024b), while the third metric was used in past ADE extraction tasks in SMM4H (Magge et al., 2021a; Weissenbacher et al., 2022a). We use precision, recall, and F_1 scores for all three evaluation metrics, where a true positive prediction means that for each tweet, the predicted annotation (either ADE ptID or ADE text span) matches the gold standard annotation. The CodaLab site for this task is `https://codalab.lisn.upsaclay.fr/competitions/18363`.

2.2 Task 2: Cross-Lingual few-shot relation extraction for pharmacovigilance in French, German, and Japanese

Task 2, like Task 1, focuses on information extraction for pharmacovigilance, but it evaluates a multilingual corpus of texts gathered from diverse sources, including patient forums, social media, and clinical reports in German, French, and Japanese. This task has two subtasks: (2a) named entity recognition (NER) for identifying mentions of drugs, disorders, and body functions, and (2b) joint NER and relation extraction (RE), evaluating both the extraction of entities and their relationships.

Dataset The training data consists of texts collected from both Twitter and a Q&A forum related to healthcare issues. It includes 392 documents in the training and 168 documents in the development set in Japanese, as well as texts in German collected from a health forum (70 documents in the training set and 23 documents in the development set). In addition, 4 French documents were added to the training set by automatically translating German documents collected from the same patient forum as German data and manually reviewed by a native French speaker. The test data comprised 118 Japanese documents, 25 German documents, and 96 French documents. Note that the translated French data did not overlap with the German data. All data were taken from the fine-grained KEEPHA corpus (Raithel et al., 2024b) and filtered for the aforementioned entity and relation types. All data were annotated with the same annotation guidelines, focusing on detecting and extracting adverse drug reactions, modeled by associating medication mentions with disorder (medical signs and symptoms) and body function mentions. The relation distribution is imbalanced (i.e., the number of "treatment_for" relations is much lower than that of "caused" relations), adding to the task's difficulty. The format of the annotations, and therefore the format of the desired predictions, is `brat` (Stenetorp et al., 2012).

Evaluation Participating systems were evaluated on CodaLab using macro precision, recall and F_1 score for both Subtask 2a and 2b across all languages in an exact match setup (only exact matching of entities are considered correct). For further analysis, single submissions were evaluated by language and with relaxed entity scores. The CodaLab site for this task is `https://codalab.lisn.upsaclay.fr/competitions/17204`.

2.3 Task 3: Multi-class classification of effects of outdoor spaces on social anxiety symptoms in Reddit

Social anxiety disorder (SAD), which is anxiety that occurs or is triggered by any situation involving other people where the individual may be judged or scrutinized, may affect up to 12% of the population at some point in their lives (Kessler et al., 2005). The onset of SAD occurs early in life, beginning by age 11 in 50% and by age 20 in 80% of patients (Stein and Stein, 2008). Individuals with SAD report experiencing symptoms for a decade before seeking treatment. However, some turn to social media platforms like Reddit to discuss their symptoms, share experiences, and seek advice for alleviating their condition. While outdoor activities in *green* -like gardens or parks- or *blue* -like rivers or lakes- outdoor spaces have been shown to benefit those with other anxiety disorders, research on their impact on SAD remains limited. To assess the perceived effects of outdoor environments on SAD, social media posts that reference these settings can be used to capture the patients' perspectives and sentiments towards them.

For this task, we challenged participants to develop a classifier to categorize posts that mention one or more words related to outdoor spaces into one of four categories: 1) positive effect, 2) neutral or no effect, 3) negative effect, or 4) unrelated, where the word mentioned is not referencing an actual outdoor space.

Dataset The data for this task was collected from the subreddit r/socialanxiety and includes only users between 12 and 25 years old (Schmidt et al., 2023). The posts from the collection were filtered to include only posts that contained at least one term from a list of about 80 keywords related to green spaces, blue spaces, and activities that take place in these spaces (e.g., running, baseball). Two annotators annotated these posts, categorizing each as nature-related or unrelated to nature. The nature-related posts were then categorized into one of the three effect categories following detailed annotation guidelines. For the subset of posts that were double annotated (n = 650), the inter-annotation agreement was k=0.796 for the initial binary annotation and k=0.72 for the multi-class annotation. The training, validation, and test sets contain 1800, 600, and 600 posts, respectively. The distribution of the classes is unbalanced, with 1,757 posts (58.6%) unrelated to nature, 298 posts (10%) reporting a positive effect attributed to the outdoor space, 214 reporting a negative effect (7%), and 731 (24.4%) labeled neutral. To prevent manual annotations during evaluation, an additional 600 decoy posts were included in the unlabeled test set provided to participants. The predictions for these decoy posts will not be evaluated.

Evaluation Participating systems were evaluated using the macro-averaged F_1 score for multi-class classification. The CodaLab site for this task is `https://codalab.lisn.upsaclay.fr/competitions/18305`.

2.4 Task 4: Extraction of the clinical and social impacts of nonmedical substance use from Reddit

Nonmedical opioid use, whether prescribed or illicit, has become a significant public health concern, leading to addiction, overdose, and associated health issues. Understanding the clinical and social impacts of nonmedical opioid use is essential for improving the treatment of opioid use disorder. It helps healthcare professionals develop more effective interventions and medications to address addiction. By studying these impacts, researchers can develop more effective prevention and education programs to reduce the occurrence of opioid abuse and its associated clinical and social consequences.

Dataset In this task, we focused on extracting two entity types from a social media dataset for analyzing clinical and social effects of substance use (Ge et al., 2024), which belonged to the category with the least number of samples: nonmedical use clinical impacts and nonmedical use social impacts. Instances in the "nonmedical use clinical impacts" category describe the clinical effects, consequences, or impacts of substance abuse or medication misuse on an individual's health, physical condition, or mental well-being. Instances in the category of "nonmedical use social impacts" describe the societal, interpersonal, or community-level effects, consequences, or impacts of substance abuse or medication misuse. These impacts may include social relationships, community dynamics, or broader social issues. The training, validation, and test sets contain 843, 259, and 278 posts, respectively. Around 27.8% of posts contain words or phrases marked as clinical or social impacts. Systems designed for this task should automatically distinguish between clinical impacts and social impacts in text data derived from Reddit, with specific spans.

Evaluation We used both token-level F_1vscore, entity-level strict F_1 score and entity-level relaxed F_1 score for evaluation. For entity-level relaxed F_1 score, we focused on the partial match between predictions and golden annotations and used the scripts from SemEval[3] to compute the score. The CodaLab site for this task is: https://codalab.lisn.upsaclay.fr/competitions/16648

2.5 Task 5: Binary classification of English tweets reporting children's medical disorders

Many children are diagnosed with disorders that can impact their daily lives and can last throughout their lifetime. For example, in the United States, 17% of children are diagnosed with a developmental disability (Zablotsky et al., 2019), and 8% of children are diagnosed with asthma (Zahran et al., 2018). Data sources for assessing the potential association of these outcomes with pregnancy exposures remain limited. Among users who reported their pregnancy on Twitter (Klein et al., 2023b), this binary classification task involved automatically distinguishing tweets that reported having a child with attention-deficit/hyperactivity disorder (ADHD), autism spectrum disorders (ASD), delayed speech, or asthma, from tweets that merely mentioned the disease. The technologies developed under this task could enable the potential use of Twitter not only for epidemiologic studies (Golder et al., 2019; Klein et al., 2022a,b, 2023a), but, more generally, to explore parents' experiences and directly target support interventions.

Dataset The training, validation, and test sets contained 7398 tweets, 389 tweets, and 1947 tweets, respectively: 3019 (31%) that reported having a child with a disorder and 6715 (69%) that did not. (Klein et al., 2024a). Inter-annotator agreement (Fleiss' kappa), based on 1000 tweets that were annotated by all three annotators, was 0.88.

Evaluation The evaluation metric for this task was the F_1 score for the class of tweets that reported having a child with a disorder. The CodaLab site for this task is: https://codalab.lisn.upsaclay.fr/competitions/17310.

2.6 Task 6: Self-reported exact age classification with cross-platform evaluation in English

Advancing the utility of social media data for research applications requires methods for automatically detecting demographic information, such as users' age, within social media study populations. Automatically identifying the exact self-reported age of social media users, rather than their age groups (the standard approach), enables large-scale use of social media data for applications that do not fit predefined age groupings of existing models. This can be particularly useful for linking specific age-related risk factors in observational studies. In this task, we focused on automatically extracting self-reported ages in posts of two social media platforms: Twitter (now X) and Reddit.

Dataset The training data consisted of 8800 labeled tweets (32% with a reported age and 68% without) (Weissenbacher et al., 2022b) and 100,000 unlabeled Reddit posts from the subreddit *r/AskDocs* that included 2-digit numbers. The validation data consisted of 2200 tweets (32% with a reported age, 68% without) and 2000 Reddit posts (53% with a reported age, 43% without) (Weissenbacher et al., 2022b). The testing data consisted

[3] https://github.com/davidsbatista/NER-Evaluation

of 2200 tweets (35% with a reported age, 65% without) and 6000 Reddit posts (60% with a reported age, 40% without). The Reddit posts were a combination of posts from two different sources, *r/socialanxiety* and *r/Dryeyes*, and the posts from the *r/socialanxiety* subreddit included only posts with reported ages on the 13 to 25 range. The inter-annotation agreement for the tweets yielded a Fleiss's kappa of 0.80, while the Reddit posts had a Cohen's Kappa inter-annotator agreement of 0.939 for the dry eye posts and 0.967 for the social anxiety posts.

Evaluation The evaluation metric was the F_1 score for the class of tweets/posts that contained the user's self-reported age. The CodaLab site for this task is: https://codalab.lisn.upsaclay.fr/competitions/17452.

2.7 Task 7: Identification of LLM or human domain-expert data annotations in the context of health-related applications.

The current widespread adoption of LLMs, like ChatGPT, for data annotation tasks has the NLP field at odds. While some researchers are embracing it due to their performance in many types of annotation tasks and certain domains, others are skeptical due to the potential underlying biases and the 'hallucinations' commonly reported with the models. It will become of paramount importance to be able to identify data annotated by LLMs and distinguish it from data annotated by humans. In this task, we provide one dataset of Tweets in Latin American Spanish containing COVID-19 symptoms with labels of annotation being made by human domain experts and by an LLM (GPT-4).

Dataset We augmented the domain-expert annotated dataset used in Task 3 of SMM4H 2023 with an equally sized dataset, consisting of non-overlapping tweets, annotated using GPT-4 with some prompt engineering. The total size of the dataset is 10,150, which was split into 4,603 tweets for training, 3,437 for validation, and 2,110 for testing.

Evaluation The evaluation metric for this task is the classification accuracy of our 'human' and 'machine' labels. The CodaLab site for this task is: https://codalab.lisn.upsaclay.fr/competitions/17405

3 Results

3.1 Task 1

Of the 27 teams registered for Task 1, 11 submitted system predictions to the CodaLab server, and 9 submitted system description papers. Table 1 shows the performance of the best submission from each team compared to the baseline system, DeepADEMiner (Magge et al., 2021b). DeepADEMiner uses a pipeline approach with BERT-based models for three sequential subtasks: 1) ADE classification, which is a binary classifier to identify whether a tweet contains ADEs; 2) ADE extraction, a sequence labeling classifier to detect ADE text spans; and 3) ADE normalization, a multi-class classifier to map the extracted ADEs to their corresponding ptIDs from MedDRA. Among all the participating teams, three teams (TLab, LHS712 and RIGA) followed the same three-step pipeline strategy, while the remaining teams only performed ADE extraction and normalization. Only two teams (SRCB and zongxiong) outperformed the baseline for ADE extraction (F_1-NER) and normalization on the overall ADEs (F_1-Norm). Additionally, three teams (SRCB, zongxiong, and TLab) outperformed the baseline for normalization on unseen ADEs (F_1-unseen).

For the ADE classification task, all three teams (Tlab, LHS712, and RIGA) fine-tuned BERT-style binary classifiers to detect the presence of ADEs in tweets. For the ADE extraction task, six teams (SRCB, Yseop, RIGA, BIT@UA, ADE Oracle, and PolyUCBS) used the BIO tagging schema and fine-tuned BERT-style language models for token classification. Meanwhile, three teams (Tlab, LHS712, and HBUT) applied prompt tuning and LLMs to directly generate ADE text spans. In the ADE normalization task, two teams (LHS712 and BIT@UA) fine-tuned a BERT-style model and a random forest multi-class classifier, respectively, to map extracted ADEs to MedDRA ptIDs. Five teams (Yseop, RIGA, HBUT, ADE Oracle, and PolyUCBS) used vector space models (VSMs) and semantic search to identify the most similar MedDRA term for each ADE. Additionally, two teams (SRCB and Tlab) employed a generate-and-rank framework, initially using VSMs to find candidate MedDRA terms and then a ranker to select the most similar term.

A total of 6 teams applyed LLMs for this task. SRCB, achieving the highest F_1-Norm of 0.536, applied LLMs for data augmentation. Specifically, the team prompted two different LLMs (GLM-4 and GPT-3.5) to rewrite ADE mentions and tweet contexts, generate synthetic tweets with diverse ADE expressions, rewrite tweets to avoid informal grammar, and generate explanations for MedDRA terms during the ranker step. They further trained an ensemble of multiple BERT-style LMs on the combination of original and augmented data for ADE extraction and normalization and showed that both two tasks benefited from the augmented data. The remaining 5 teams mainly leverage the in-context learning of LLMs for the extraction and normalization task. For instance, Tlab team used retrieved tweets as few-shot examples as input to Llama 3 to extract ADEs, and a retrieval augmented generation framework to take the candidate MedDRA terms and the original tweet as input to GPT4, and generate the best MedDRA candidate term. Similarly, the LHS712 team also used a one-shot example as input and experimented with 10 different prompts for ADE extraction, but they achieved worse performance. The RIGA team used GPT4 with prompt tuning to find potential ADE text spans, which, along with the original tweets, were then fed as input for classification and extraction.

In conclusion, Task 1 underscored the challenges existing pipeline systems face in ADE extraction and normalization tasks. While LLMs have been widely adopted for these tasks, using them directly as in-context few-shot learners yielded poorer performance than BERT-style models. However, leveraging LLMs for data augmentation can enhance the performance of BERT-style models.

3.2 Task 2

Out of the 13 teams registered for Task 2, only 3 teams submitted system predictions to the CodaLab server, and only 2 teams submitted system description papers. Furthermore, these two teams did not participate in all subtasks or cover all languages. The results for Task 2a (NER) and Task 2b (joint NER and RE) in detecting Adverse Drug Reactions (ADRs) across German, French, and Japanese are displayed in Table 2.

Our baseline system first employed the Pytorch-IE framework (Binder et al., 2024)to predict entities with BERT-style token classification models. For the German data, we fine-tuned a German version of BERT (Devlin et al., 2019)); for the Japanese data, we fine-tuned the multilingual XLM-RoBERTa (Conneau et al., 2019)); and for the French data, we reused the model fine-tuned on Japanese without additional fine-tuning. The predicted entities from these models were then used as input for prompt templates to generate the relationships between the detected entities. These prompts included examples from the training data, brief definitions of the desired entities and relationships, and a requirement for the model to explain its decisions. The model we used for prompting was the open Llama-3-8B-UltraMedical (Zhang et al., 2024).

Task 2a: Team Yseop participated in the NER task for Japanese and French, while team HBUT submitted predictions for all three languages.

For French NER, Team Yseop used a combination of advanced language models, including the large language models Mistral-7B (Jiang et al., 2023) and DrBERT-CASM2 (Labrak et al., 2023), along with the medkit framework. For Japanese NER, they employed a Japanese-Multilingual Dictionary (JMdict) and a Japanese medical language model based on RoBERTa (Liu et al., 2019), pre-trained on Japanese case reports and fine-tuned for NER using MedTxt-CR (Yada et al., 2022). They achieved a macro F_1 score of 48.92 across the three languages (including German, for which they did not provide predictions).

Team HBUT focused solely on the NER task for all three languages. Their methodology also employed LLMs. They explored three distinct prompting strategies to identify the most effective approach for NER. After evaluating two different LLMs, they selected GLM-3-Turbo (Zeng et al., 2023) as their preferred model. To adapt the task for LLMs, they designed specific prompts to obtain structured output. These outputs were then post-processed into the desired `brat` format. Evaluating these predictions against the gold entities resulted in an F_1 score of 36.9 for Team HBUT.

Task 2b: For the Japanese Relation Extraction task, Team Yseop reused the Japanese XLM-RoBERTa model and fine-tuned it with the provided training data. Since Team Yseop was the only team to submit predictions for the RE task, they were automatically declared the winners of the challenge, achieving an F_1 score of 1.89.

Summary: Table 2 presents the best-performing system from each team. Notably, neither team surpassed the baseline in any of the tasks, but Team HBUT achieved higher precision in the NER task compared to the other team and the baseline. Ex-

Team	F_1-Norm	F_1-NER	F_1-Unseen	System Summary
SRCB	**0.536**	**0.521**	**0.494**	Ensemble of BERT-style models for extraction and normalization, LLM-based data augmentation
zongxiong	0.528	0.513	0.492	–
Baseline	0.439	0.481	0.323	BERT-style models in 3-step pipeline system
Yseop	0.400	0.472	0.295	BERT-style model for extraction, BERT-style VSM for normalization
TLab	0.359	0.392	0.363	BERT-style model for classification, Llama3 for extraction, GPT-4 for normalization
LHS712	0.354	0.338	0.259	BERT-style models for classification and normalization, GPT-4 for extraction, BioBERT for normalization
RIGA	0.318	0.403	0.212	GPT-4 for preprocessing, and BERT-style models for classification and extraction, OpenAI Embeddings as VSM for normalization
BIT@UA	0.295	0.397	0	RoBERTa-base for span extraction, random forest classifier for normalization
Proddis	0.221	0.001	0.098	–
HBUT	0.205	0.216	0.106	GLM for extraction, ensemble of BERT-style models for normalization
ADE Oracle	0.082	0.132	0.014	Spacy tool for extraction, BERT-style VSM for normalization
PolyuCBS	0.044	0.010	0	LLM for preprocessing ,BERT-style model for extraction, SapBERT for normalization

Table 1: System summaries and micro-averaged F_1-scores for **task 1**. F_1*-Norm* is the ADE normalization score for all ptIDs, F_1*-NER* is the ADE extraction score, and F_1*-Unseen* is the ADE normalization score for unseen ptIDs. '-' in the *System Summary* indicates no system description paper was submitted.

Team	Task	Languages	P	R	F_1	System Summary
Yseop	NER	fr, ja	58.31	42.14	48.92	fr: Mistral-7B + DrBERT-CASM2 + medkit; ja: dictionary + RoBERTa
HBUT	NER	de, fr, ja	**60.52**	26.54	36.90	GLM-3-Turbo prompting + post-processing
baseline	NER	de, fr, ja	47.55	**58.83**	**52.60**	BERT, XLM-RoBERTa
Yseop	RE	ja	02.24	01.63	01.89	ja: RoBERTa fine-tuned
baseline	RE	de, fr, ja	**04.25**	**06.81**	**05.23**	language-specific prompting with Llama-3-8B-UltraMedical

Table 2: Summary of the submitted systems for **Task 2**. Subtask 2a: Named Entity Recognition (NER). Subtask 2b: joint NER and Relation Extraction (RE). The scores show exact macro F_1 score ($\mathbf{F_1}$), precision (**P**), and recall (**R**) as reported in CodaLab. The underlined scores belong to the winning systems/teams of the challenge.

cluding the baseline, Team Yseop won both subtasks in Task 2. For a more detailed description of the task and its results, we refer the reader to Raithel et al. (2024a).

In conclusion, participants employed diverse methods such as dictionary-based approaches, transformers, and LLMs, yet the task remains highly challenging. This difficulty likely stems from the task's multi-language nature, the presence of noisy, user-generated texts, and the limited number of training instances. Challenges also included generating correct output formats, identifying valid entity spans, and establishing accurate relations (some predictions included relations to non-existent entities). Team Yseop's approach, which combined outputs from multiple models in an ensemble manner, appears promising. Understanding how Team HBUT achieved high precision for entities would be particularly insightful. Combining their approach with our baseline prompting strategy could potentially enhance overall performance. Despite the use of LLMs, none of the participating teams fully solved the task of joint multilingual named entity recognition and relation extraction. This underscores the need for further exploration and possible integration of different methodologies.

3.3 Task 3

Of 37 teams registered for Task 3, 15 submitted system predictions, and 12 submitted task description papers. Table 3 displays the macro-average F1 score, precision, and recall for the top-performing submission from each team. The top 5 teams achieved closely matched scores, with only a 0.05 difference between the first and fifth positions. Team CTYUN-AI attained the high-

Team	F_1	P	R	System Summary
CTYUN-AI	**0.692**	**0.704**	**0.686**	Qwen-72B-Chat, data augmentation
1024m	0.679	0.677	0.682	BART-Large (2-stage)
PCIC	0.655	0.687	0.636	XLNet-large, data augmentation (traditional + paraphrasing (T5-base))
Dilab	0.654	0.654	0.661	RoBERTa-Large with Fuzzy string matching
Dolomites	0.642	0.67	0.623	Mistral-7B, fine-tuning (QLoRA), Multi-Task Learning Data Augmentation (In-domain + Drills)
AAST-NLP	0.635	0.631	0.644	RoBERTa-Large
ThangDLU	0.627	0.62	0.644	BART-base
Golden_Duck	0.596	0.603	0.601	RoBERTa-base: Concatenation of Mean pooling, CLS, and Attention Head
IMS_medicALY	0.563	0.629	0.534	SocBERT
LAMA	0.545	0.633	0.536	Pipeline: MentalBERT (binary classification: related or unrelated) + RoBERTa (multiclass: pos, neg or neu)
gcortal	0.414	0.425	0.418	–
Transformers	0.413	0.431	0.52	RoBERTa-Large, under-sampling
interrupt-driven	0.358	0.365	0.411	Relevance-weighted sentiment analysis model, Passive-Aggressive Regressor
Omkar_Khade	0.233	0.683	0.287	–
TeamZSA_codalab	0.196	0.167	0.27	–

Table 3: System summaries and macro-averaged F_1-score (F_1), precision (P), and recall (R) for **Task 3**: multi-class classification of effects of outdoor spaces on social anxiety symptoms in Reddit.

Team Name	Relaxed F_1	Strict F_1	Token-level F_1	System Summary
UKYNLP	**0.462**	0.171	**0.531**	Span-based encoder-only model leveraging BERT and ALBERT, combined with a BiLSTM layer for classification of entity types.
Dolomites	0.448	**0.208**	0.496	Experimented with two MTL-DA techniques to fine-tune Mistral (7B) with QLoRA for low-resource settings
LHS712 NV	0.314	0.008	0.052	Fine-tuning pre-trained BERT

Table 4: Brief system approaches and evaluation metric results for **Task 4**: Extraction of the clinical and social impacts of nonmedical substance use from Reddit.

est macro-average F1 score (0.692) by employing the LLM Qwen-72b-Chat, which was pre-trained and fine-tuned for classification. They also implemented data augmentation to balance classes, involving the random shuffling of strings within the original post based on delimiters.

The team with the second-highest macro-average F1 score (0.679), 1024m, built its system around a BART-large model using a two-stage strategy: initially, posts are classified as either class 0 or not, followed by categorizing non-0 posts into one of the remaining three classes. The third-highest scoring team, PCIC, achieved a macro-average F1 score of 0.655 by employing XLNet-large. Their approach included using a combined loss function, implementing data augmentation to balance class distributions, and increasing the token window size to 256.

Only two teams, CTYUN-AI and Dolomites, achieved their highest performance using an LLM. Dolomites (F1 score: 0.642) utilized Mistral-7B and employed multi-task learning data augmentation (MTL-DA) techniques to fine-tune the LLM. These techniques included in-domain augmentation from similar resources such as Reddit, as well as drills that decomposed the task into smaller subtasks to generate replicated data. This approach outperformed GPT-4 with zero-shot or few-shot learning on the validation set. The remaining teams achieved their best results using transformer-based models such as BART, RoBERTa, and XLNet.

In conclusion, Task 3 underscored the effectiveness of transformer models in classification tasks, with only a limited number of teams effectively leveraging LLMs.

3.4 Task 4

Out of 23 teams registered for Task 4, only 3 teams ultimately submitted task description papers. Table 4 presents the performances of three different teams. UKYNLP achieved the highest scores, with a *Relaxed* F_1 score of 0.462 and a *Token-level* F_1 score of 0.531, demonstrating excellence in both broad recognition and fine-grained token-level accuracy. UKYNLP experimented with both BERT and ALBERT models combined with a BiLSTM

Team	F_1	P	R	System Summary
CTYUN-AI	**0.956**	**0.954**	0.959	Qwen-72B (fine-tuned), multi-task learning (Task 6)
LT4SG	0.938	0.930	0.946	BERTweet-Large ensemble
PolyuCBS	0.935	0.954	0.917	Llama-2-7B (fine-tuned), LoRA
UTRad-NLP	0.933	0.932	0.934	DeBERTa-V3-Large, GPT-4 data augmentation
Golden_Duck	0.928	0.919	0.937	RoBERTa-Large
Chaai	0.927	0.907	0.949	Twitter-RoBERTa, GPT-4 (zero-shot), under-sampling
Baseline	0.927	0.923	0.930	RoBERTa-Large
UKYNLP	0.924	0.924	0.924	DeBERTa-V3-Large
KUL	0.923	0.906	0.940	BERTweet, data augmentation, R-Drop
1024m	0.918	0.923	0.912	BART-Large
SMC	0.901	0.885	0.917	MentalBERT, PsychBERT, TwHIN-BERT, DistilBERT, data augmentation
Transformers	0.900	0.854	0.950	RoBERTa-Large, under-sampling
DILAB	0.898	0.883	0.914	Twitter-RoBERTa-Base-Sentiment ensemble
HALELab-NITK	0.868	0.858	0.879	RoBERTa-Base
Thang-DLU	0.841	0.844	0.839	T5-Small, GPT data augmentation
BIT@UA	0.840	0.829	0.851	BERT-Base
PhoenixTrio_918	0.823	0.721	0.959	RoBERTa-Large, 5-fold cross-validation
MUET	0.671	0.508	**0.988**	-
HULAT-UC3M	0.633	0.702	0.576	-
Be Better Health	0.522	0.630	0.445	-
Z-AGI Labs	0.357	0.321	0.401	-

Table 5: System summaries and F_1-score (F_1), precision (P), and recall (R) for **Task 5**: Binary classification of English tweets reporting children's medical disorders.

layer for NER. Their BERT model achieved the highest F_1 score of 0.462 on the test set, surpassing the ALBERT model with BiLSTM.

Dolomites utilized multi-task learning and data augmentation techniques, achieving moderate success overall with a strict F_1 score of 0.208. This suggests their approach is effective for precise entity matching. They employed a strategy of extracting smaller tasks (drills) from the target dataset and replicating the training set to include additional examples for these drills.

In contrast, team LHS712 NV showed the lowest performance across all metrics, indicating potential challenges in model implementation or task-specific tuning despite utilizing robust BERT-style models.

3.5 Task 5

Out of 48 teams registered for Task 5, 20 teams submitted system predictions to the Codalab server, and 16 teams ultimately submitted task description papers. Table 5 presents the F_1, precision, and recall scores for a RoBERTa-large baseline classifier (Klein et al., 2024a) and the best-performing system from each of the 20 teams for Task 5. CTYUN-AI achieved the highest F_1 score (0.956) and precision (0.954) using a Qwen-72B LLM and multi-task learning, improving upon the baseline (0.927) by approximately 0.03. Initially, they continued pre-training a Qwen-72B LLM using unlabeled Reddit posts from Task 6. Subsequently, they fine-tuned two additional Qwen-72B LLMs using labeled tweets and Reddit posts from Task 6, deploying them for binary classification of the unlabeled Reddit posts. They further refined this LLM by focusing on Reddit posts where the two classifiers agreed, combining them with labeled tweets and Task 6 Reddit posts for additional fine-tuning. Finally, they fine-tuned this LLM further using labeled tweets from Task 5. An ablation study is necessary to determine the precise impact of LLMs and multi-task learning on their performance.

Among the other top-performing teams that exceeded the baseline (0.927), PolyuCBS achieved an F_1 score of 0.935 using a Llama-2-7B LLM fine-tuned with LoRA, narrowly edged out by LT4SG's ensemble of BERTweet-Large models with an F_1 score of 0.938. Of these teams, only CTYUN-AI and PolyuCBS utilized LLMs in their approaches. Additionally, two other top teams used a GPT-4 LLM to augment training data for a DeBERTa-V3-Large classifier (UTRad-NLP) and validate predictions from a Twitter-RoBERTa classifier (Chaii). All teams that submitted system descriptions employed deep neural network architectures based on pre-trained transformer models.

Team	F_1	P	R	System Summary
CTYUN-AI	**0.970**	**0.976**	0.963	fine-tuned Qwen-72B-Chat, data augmentation by random shuffling, ensemble labeling of unlabeled Reddit data and cross-task training
1024m	0.959	0.953	**0.965**	BART-Large
Dolomites	0.957	0.965	0.949	Mistral-7B, fine-tuning (QLoRA), Multi-Task Learning Data Augmentation (In-domain + Drills)
AAST-NLP	0.946	0.932	0.959	Ensemble of BERTweet, RoBERTa and Mistral-7b, rule-based data augmentation from unlabeled data
UTRad-NLP	0.936	0.947	0.926	DeBERTa-V3-Large, synthetic data augmentation with GPT-4
Baseline	0.900	0.902	0.897	RoBERTa-Large
SMM4H-TIET	0.900	0.916	0.884	BERTweet, back-translation augmentation of minority class, under-sampling of majority class
IITRoorkee	0.878	0.899	0.858	RoBERTa

Table 6: System summaries and F_1-scores (F_1), precision (P), and recall (R) for **Task 6**: Self-reported exact age classification with cross-platform evaluation in English.

Team Name	Accuracy	System Summary
Baseline	0.82	Fine-tuned COVID-Twitter-BERT w 500K silver-standard annotations
712forTask7	0.5166	Fine-tuned BETO
BrainStorm	0.5090	Fine-tuned two different BERT-style models, topical embeddings from BERTopic
Deloitte	0.5109	Zero-shot prompt tuning on GPT4

Table 7: System summaries and classification accuracy results for **Task 7**.

3.6 Task 6

Out of 27 teams registered for Task 6, 7 teams ultimately submitted task description papers. Table 6 displays the F_1 score, precision, and recall for the RoBERTa-Large baseline classifier (Klein et al., 2022c) and the top-performing system runs of these 7 teams. Among these teams, four utilized LLMs: CTYUN-AI applied Qwen-72B-Chat, Dolomites used Mistral-7B, and AAST-NLP integrated Mistral-7B into their ensemble models. Additionally, UTRad-NLP employed synthetic data generated with GPT-4 to augment their training dataset. The remaining teams focused on transformer models: team 1024m utilized BART-Large, team UTRad-NLP employed DeBERTa-V3-Large, team AAST-NLP used an ensemble of BERTweet, RoBERTa, and Mistral-7B, team SMM4H-TIET used BERTweet, and team IITRoorkee used RoBERTa.

The top-performing team, CTYUN-AI, achieved the highest F_1 score (0.970) and precision (0.976) by fine-tuning the Qwen-72B-Chat model. Their approach included data augmentation through random sentence shuffling, ensemble labeling of the unlabeled Reddit data provided, and cross-task training, significantly enhancing their model's performance. Only two teams employed different approaches to label the provided unlabeled Reddit posts: CTYUN-AI used an ensemble model, and AAST-NLP employed a rule-based approach. Notably, team 1024m achieved the highest recall, utilizing a BERT-style model.

Comparing the use of LLMs against traditional BERT-style models, LLMs generally demonstrated superior performance. For example, CTYUN-AI's use of Qwen-72B-Chat outperformed the baseline RoBERTa-Large model by 0.070 in terms of F_1 score. This trend was consistent across other teams' results, where LLMs, when fine-tuned and augmented with advanced techniques, consistently achieved higher precision and recall compared to traditional BERT-style models. However, BERT-style models also performed well, especially when used in ensembles or enhanced with data augmentation strategies.

3.7 Task 7

Out of 19 teams registered for Task 7, only 3 teams ultimately submitted system predictions on the test set and system description papers. Table 7 pdisplays the classification accuracy for the best system run from these 3 teams on the unseen test set. The baseline system utilized a fine-tuned COVID-Twitter-BERT (Müller et al., 2023) trained on the provided dataset and an additional 500K 'silver-standard' tweets sourced from Banda et al. (2021). These tweets were generated using weak supervision annotation for half and GPT-4 for the other half. We hypothesize that this extensive data augmentation played a significant role in the performance disparity between the participant scores and the baseline score.

Among the participants, two teams focused on traditional BERT-style models. One team employed BETO, a BERT-based model for Spanish text (Cañete et al., 2020) , while another team translated Spanish-language tweets into English and utilized BERTopic embeddings (Grootendorst, 2022). The third team, Deloitte, used GPT-4 to classify the provided tweets.

One team, 712forTask7, conducted an analysis of the training dataset, highlighting minimal differences among the tweets, which posed challenges for traditional approaches. The substantial data augmentation in the baseline system likely contributed significantly to its performance advantage. It would be intriguing to explore whether participant teams could achieve comparable performance using the same augmented dataset with their respective approaches.

4 Conclusion

This paper provides an overview of the SMM4H 2024 shared tasks. This year, seven tasks were proposed, reflecting the growing interest and participation in the SMM4H shared tasks. The top-performing teams predominantly used transformer-based models, including encoder-based LMs like DeBERTa-v3-large and decoder-based LLMs like GPT-4 or Qwen-72B-Chat. These teams frequently employed LLM-based data augmentation techniques to address issues such as data imbalance, unseen examples, and domain mismatch in social media data. Notably, out of 38 teams that submitted a system description paper, 11 participated in multiple tasks, often using the same systems fine-tuned for different tasks. This trend signifies an important effort within the community to develop high-performing classifiers and label sequencers that are both generalizable and reusable.

Acknowledgments

The work for SMM4H 2024 at Cedars-Sinai Medical Center was supported by the National Institutes of Health (NIH) National Library of Medicine (NLM) [grant number R01LM011176]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

The work of the organizers of Task 2 was supported by the Cross-ministerial Strategic Innovation Promotion Program (SIP) on "Integrated Health Care System" Grant Number JPJ012425, by JPMJCR20G9, ANR-20-IADJ-0005-01, and DFG-442445488 under the trilateral ANR-DFG-JST AI Research project KEEPHA, and by the German Federal Ministry of Education and Research under the grant BIFOLD24B.

Creation of the dataset for task 4 was supported in part by the National Institute on Drug Abuse (NIDA) of the National Institutes of Health (NIH) under award number R01DA057599. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

The work of the organizer of Task 7 was supported by a Google Award for Inclusion Research (AIR).

The authors thank those who contributed to annotating the data, the program committee of the #SMM4H 2024 Workshop, and additional peer reviewers of the system description papers.

References

Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Ekaterina Artemova, Elena Tutubalina, and Gerardo Chowell. 2021. A large-scale covid-19 twitter chatter dataset for open scientific research—an international collaboration. *Epidemiologia*, 2(3):315–324.

Arne Binder, Leonhard Hennig, and Christoph Alt. 2024. Pytorch-ie: Fast and reproducible prototyping for information extraction. *Preprint*, arXiv:2406.00007.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Pew Research Center. SM Fact Sheet 2021. https://www.pewresearch.org/internet/fact-sheet/social-media/. Accessed 2nd August 2023.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

N. Elliott and G. Sverdlov. 2012. Global SM adoption. the social marketing playbook: Master the next wave of social.

Yao Ge, Sudeshna Das, Karen O'Connor, Mohammed Ali Al-Garadi, Graciela Gonzalez-Hernandez, and Abeed Sarker. 2024. Reddit-impacts: A named entity recognition dataset for analyzing clinical and social effects of substance use derived from social media. *Preprint*, arXiv:2405.06145.

S. Golder, S. Chiuve, D. Weissenbacher, A. Klein, K. O'Connor, M. Bland, M. Malin, M. Bhattacharya, L. J. Scarazzini, and G. Gonzalez-Hernandez. 2019. Pharmacoepidemiologic evaluation of birth defects from health-related postings in social media during pregnancy. *Drug Saf*, 42(3):389–400.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Ronald C. Kessler, Wai Tat Chiu, Olga Demler, and Ellen E. Walters. 2005. Prevalence, Severity, and Comorbidity of 12-Month DSM-IV Disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, 62(6):617–627.

A. Z. Klein, J. A. Gutiérrez Gómez, L. D. Levine, and G. Gonzalez-Hernandez. 2024a. Using Longitudinal Twitter Data for Digital Epidemiology of Childhood Health Outcomes: An Annotated Data Set and Deep Neural Network Classifiers. *J Med Internet Res*, 26:e50652.

A. Z. Klein, S. Kunatharaju, S. Golder, L. D. Levine, J. C. Figueiredo, and G. Gonzalez-Hernandez. 2023a. Association between COVID-19 during pregnancy and preterm birth by trimester of infection: a retrospective cohort study using longitudinal social media data. *medRxiv*.

A. Z. Klein, S. Kunatharaju, K. O'Connor, and G. Gonzalez-Hernandez. 2023b. Pregex: Rule-Based Detection and Extraction of Twitter Data in Pregnancy. *J Med Internet Res*, 25:e40569.

A. Z. Klein, K. O'Connor, and G. Gonzalez-Hernandez. 2022a. Toward using Twitter data to monitor COVID-19 vaccine safety in pregnancy: proof-of-concept study of cohort identification. *JMIR Form Res*, 6(1):e33792.

A. Z. Klein, K. O'Connor, L. D. Levine, and G. Gonzalez-Hernandez. 2022b. Using Twitter data for cohort studies of drug safety in pregnancy: proof-of-concept with B-blockers. *JMIR Form Res*, 6(6):e36771.

Ari Z Klein, Juan M Banda, Yuting Guo, Ana Lucia Schmidt, Dongfang Xu, Ivan Flores Amaro, Raul Rodriguez-Esteban, Abeed Sarker, and Graciela Gonzalez-Hernandez. 2024b. Overview of the 8th Social Media Mining for Health Applications (SMM4H) shared tasks at the AMIA 2023 Annual Symposium. *Journal of the American Medical Informatics Association*, 31(4):991–996.

Ari Z Klein, Arjun Magge, and Graciela Gonzalez-Hernandez. 2022c. Reportage: Automatically extracting the exact age of twitter users based on self-reports in tweets. *PloS one*, 17(1):e0262087.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickaël Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. Drbert: A robust pre-trained model in french for biomedical and clinical domains. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16207–16221.

Robert Leaman, Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jian Yang, and Graciela Gonzalez. 2010. Towards Internet-age pharmacovigilance: Extracting adverse drug reactions from user posts in health-related social networks. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 117–125, Uppsala, Sweden. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692 [cs]*.

Arjun Magge, Ari Klein, Antonio Miranda-Escalada, Mohammed Ali Al-Garadi, Ilseyar Alimova, Zulfat Miftahutdinov, Eulalia Farre, Salvador Lima López, Ivan Flores, Karen O'Connor, Davy Weissenbacher, Elena Tutubalina, Abeed Sarker, Juan Banda, Martin Krallinger, and Graciela Gonzalez-Hernandez. 2021a. Overview of the sixth social media mining for health applications (#SMM4H) shared tasks at NAACL 2021. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 21–32, Mexico City, Mexico. Association for Computational Linguistics.

Arjun Magge, Elena Tutubalina, Zulfat Miftahutdinov, Ilseyar Alimova, Anne Dirkson, Suzan Verberne, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021b. DeepADEMiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter. *Journal of the American Medical Informatics Association*, 28(10):2184–2192.

Martin Müller, Marcel Salathé, and Per E. Kummervold. 2023. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *Frontiers in Artificial Intelligence*, 6.

Lisa Raithel, Philippe Thomas, Bhuvanesh Verma, Roland Roller, Hui-Syuan Yeh, Shuntaro Yada, Cyril Grouin, Shoko Wakamiya, Eiji Aramaki, Sebastian Möller, and Pierre Zweigenbaum. 2024a. Overview of #SMM4H 2024 – Task 2: Cross-lingual few-shot relation extraction for pharmacovigilance in french, german, and japanese. In *Proceedings of The Ninth Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, Bangkok, Thailand. Association for Computational Linguistics.

Lisa Raithel, Hui-Syuan Yeh, Shuntaro Yada, Cyril Grouin, Thomas Lavergne, Aurélie Névéol, Patrick Paroubek, Philippe Thomas, Tomohiro Nishiyama, Sebastian Möller, Eiji Aramaki, Yuji Matsumoto, Roland Roller, and Pierre Zweigenbaum. 2024b. A dataset for pharmacovigilance in German, French, and Japanese: Annotating adverse drug reactions across languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 395–414, Torino, Italia. ELRA and ICCL.

Ana Lucia Schmidt, Karen O'Connor, Graciela Gonzalez Hernandez, and Raul Rodriguez-Esteban. 2023. Studying social anxiety without triggering it: Establishing an age-controlled cohort of social media users for observational studies. *medRxiv*.

Murray B. Stein and Dan J. Stein. 2008. Social anxiety disorder. *Lancet*, 371(9618):1115–1125. London, England.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: A Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Andrea C Tricco, Wasifa Zarin, Erin Lillie, Serena Jeblee, Rachel Warren, Paul A Khan, Reid Robson, Ba' Pham, Graeme Hirst, and Sharon E Straus. 2018. Utility of social media and crowd-intelligence data for pharmacovigilance: a scoping review. *BMC medical informatics and decision making*, 18:1–14.

Davy Weissenbacher, Juan Banda, Vera Davydova, Darryl Estrada Zavala, Luis Gasco Sánchez, Yao Ge, Yuting Guo, Ari Klein, Martin Krallinger, Mathias Leddin, Arjun Magge, Raul Rodriguez-Esteban, Abeed Sarker, Lucia Schmidt, Elena Tutubalina, and Graciela Gonzalez-Hernandez. 2022a. Overview of the seventh social media mining for health applications (#SMM4H) shared tasks at COLING 2022. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 221–241, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Davy Weissenbacher, Juan Banda, Vera Davydova, Darryl Estrada Zavala, Luis Gasco Sánchez, Yao Ge, Yuting Guo, Ari Klein, Martin Krallinger, Mathias Leddin, et al. 2022b. Overview of the seventh social media mining for health applications (# smm4h) shared tasks at coling 2022. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 221–241.

Shuntaro Yada, Shoko Wakamiya, Yuta Nakamura, and Eiji Aramaki. 2022. Real-MedNLP: Overview of REAL document-based MEDical Natural Language Processing Task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*, Tokyo, Japan.

B. Zablotsky, L. I. Black, M. J. Maenner, L. A. Schieve, M. L. Danielson, R. H. Bitsko, S. J. Blumberg, M. D. Kogan, and C. A. Boyle. 2019. Prevalence and Trends of Developmental Disabilities among Children in the United States: 2009-2017. *Pediatrics*, 144(4).

H. S. Zahran, C. M. Bailey, S. A. Damon, P. L. Garbe, and P. N. Breysse. 2018. Vital Signs: Asthma in Children - United States, 2001-2016. *MMWR Morb Mortal Wkly Rep*, 67(5):149–155.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: An Open Bilingual Pre-trained Model. *Preprint*, arXiv:2210.02414.

Kaiyan Zhang, Ning Ding, Biqing Qi, Sihang Zeng, Haoxin Li, Xuekai Zhu, Zhang-Ren Chen, and Bowen Zhou. 2024. Ultramedical: Building specialized generalists in biomedicine. `https://github.com/TsinghuaC3I/UltraMedical`.