Overview of the Seventh Social Media Mining for Health Applications (#SMM4H) Shared Tasks at COLING 2022

Davy Weissenbacher,[1]† Juan M. Banda,[2] Vera Davydova,[3] Darryl Estrada-Zavala,[4] Luis Gascó,[4] Yao Ge,[5] Yuting Guo,[5] Ari Z. Klein,[6] Martin Krallinger,[4] Mathias Leddin,[7] Arjun Magge,[6] Raul Rodriguez-Esteban,[7] Abeed Sarker,[5] Ana Lucía Schmidt,[7] Elena Tutubalina,[8,9] Graciela Gonzalez-Hernandez[1]

[1]Cedars-Sinai Medical Center, Los Angeles, CA, USA
[2]Georgia State University, Atlanta, GA, USA
[3]Sber AI, Moscow, Russia
[4]Barcelona Supercomputing Center (BSC), Barcelona, Spain
[5]Emory University, Atlanta, GA, USA
[6]University of Pennsylvania, Philadelphia, PA, USA
[7]Roche Innovation Center, Basel, Switzerland
[8]Kazan Federal University, Kazan, Russia
[9]AIRI, Moscow, Russia
†Corresponding author: davy.weissenbacher@cshs.org

Abstract

For the past seven years, the Social Media Mining for Health Applications (#SMM4H) shared tasks have promoted the community-driven development and evaluation of advanced natural language processing systems to detect, extract, and normalize health-related information in public, user-generated content. This seventh iteration consists of ten tasks that include English and Spanish posts on Twitter, Reddit, and WebMD. Interest in the #SMM4H shared tasks continues to grow, with 117 teams that registered and 54 teams that participated in at least one task—a 17.5% and 35% increase in registration and participation, respectively, over the last iteration. This paper provides an overview of the tasks and participants' systems. The data sets remain available upon request, and new systems can be evaluated through the post-evaluation phase on CodaLab.

1 Introduction

For the past seven years, the Social Media Mining for Health (#SMM4H) Workshop has been a competitive platform to promote the development and evaluation of advanced natural language processing systems to detect, extract, and normalize health-related information in user generated content publicly available online. For the seventh iteration of the Workshop shared tasks, researchers from international institutions challenged the community with ten tasks. The tasks run on various media (tweets, reviews and Reddit posts) and languages (English and Spanish): classification, detection and normalization of adverse events mentions in tweets (Task 1), classification of stance and premise in tweets about health mandates related to COVID-19 (Task 2), classification of changes in medication treatments in tweets and WebMD reviews (Task 3), classification of tweets self-reporting exact age (Task 4), classification of tweets containing self-reported COVID-19 symptoms - in Spanish (Task 5), classification of tweets which indicate self-reported COVID-19 vaccination status (Task 6), classification of self-reported intimate partner violence on Twitter (Task 7), classification of self-reported chronic stress on Twitter (Task 8), classification of Reddit posts self-reporting exact age (Task 9), detection of disease mentions in tweets – in Spanish (Task 10).

This iteration shows that the interest of the community for the shared tasks continues to grow with 117 teams registered, representing a 17.5% growth compare to the last iteration in 2021. The growth of interest translated in an increase of participation with 54 teams having submitted their predictions for at least one task, a growth of 35% compared to last year participation. While all task organizers were free to change the modality for their task, we recommended generic guidelines to organize the tasks. We allowed all participants to register in one or more tasks. Upon acceptance, we provided

Proceedings of the 29th International Conference on Computational Linguistic, pages 221–241
October 12–17, 2022.

the participants with a training and a validation set for each task they registered in during the practice period. We released unlabeled test sets at the beginning of the evaluation period of the competition, a period which was 4 days long. During this period, each team could upload up to 3 predictions sets for the labels of the test sets to the web-based platform, Codalab. Codalab automatically evaluated their performance. We report in this overview the results achieved with the best set of three predictions sets submitted. We left the post-evaluation period opened indefinitely on Codalab for all tasks run during this iteration. Interested readers can request the datasets and upload their own predictions to the Codalab server to compare their performance with the winners of each task.

In Section 2, we briefly describe and motivate the ten tasks of the competition. In Section 3, we present the performance and a short interpretation of the results for each task. In Appendix 4, we provide the list of the publications describing the systems of the competing teams along with the team IDs referring them in Section 3.

2 Tasks

2.1 Task 1: Classification, detection and normalization of Adverse Events (AE) mentions in tweets (in English)

For Task 1, participants were asked to develop methods to extract adverse drug effects (ADE) from tweets containing drug mentions. Such automated methods are considered necessary for social media pharmacovigilance efforts for monitoring emergence of ADEs in post-market surveillance of drugs, especially in vulnerable populations such as children, pregnant women and the elderly who are often excluded from clinical trials. Reliable signals generated from such social media pharmacovigilance pipelines can be complementary to traditional pharmacovigilance efforts such as voluntary consumer and health provider reporting. ADE mining in tweets has been the longest running tasks at the SMM4H shared task. Task 1 presents three subtasks in increasing order of complexity: (Task 1a) Identify tweets that contain one or more ADEs, (Task 1b) in addition to Task 1a, extract the text span of reported ADEs in tweets, and (Task 1c) in addition to Tasks 1a and 1b, normalize extracted spans to MedDRA ontology's preferred terms `https://meddra.org/`. For this task, we used the same dataset as the previous year. During the training state, the dataset contains a total of 18,300 tweets with 17,385 tweets for training and 915 tweets for validation (Magge et al., 2020). During the evaluation stage, we performed a blind evaluation on 10,984 tweets. The tweets were manually annotated by experts with: (a) binary labels of *hasADE* and *noADE* indicating presence of one or more ADEs, (b) starting and ending indices of the ADE mention(s) in the text, and (c) the normalized MedDRA lower-level term (LLT) that were evaluated at the higher preferred term (PT) level. MedDRA contains about 79,000 LLT terms and about 23,000 preferred terms, making it important for the developed systems to be capable of identifying and normalizing ADEs that were not occurring in the training set. Task 1 presents multiple technical challenges such as class imbalance and normalizing into a large potential label space. Submissions were evaluated and ranked based on the F_1-score for the *ADE* class for subtask 1a, F_1-score for overlapping ADE mentions for subtask 1b, and F_1-score for overlapping ADE mentions with matching preferred term IDs for subtask 1c. Task 1 is hosted on Codalab at `https://codalab.lisn.upsaclay.fr/competitions/2073`.

2.2 Task 2: Classification of stance and premise in tweets about health mandates related to COVID-19 (in English)

For Task 2, we focus on argument mining (or argumentation mining) for extracting arguments from COVID-related tweets. Tweets express views towards three claims/topics associated with governments' restrictions: (i) keeping schools closed, (ii) stay-at-home orders, and (iii) wearing a face mask. Systems for detecting people's stances and their premises about governments' health mandates related to COVID-19 can help to gauge the level of cooperation in society which is essential for stopping the spread of the virus. The task consists of two sub-tasks: (i) **Task 2a** on stance detection, and (ii) **Task 2b** on premise classification.

Task 2a: stance detection The first sub-task aims to determine the point of view (stance) of the text's author in relation to the given claim. The tweets are manually annotated for stance according to three categories: in-favor, against, and neither.

Task 2b: premise classification The second sub-task is to predict whether at least one premise/argument is mentioned. A given tweet is considered as having a premise if it contains

Set	Stance classes			Premise classes	
	in-favor	against	neither	1	0
Train	1346	874	1336	1331	2225
Valid	244	158	198	221	379
Test	526	570	904	716	1284

Table 1: Statistics of the Tasks 2a (stance) & 2b (premise) datasets

a statement that can be used as an argument in a discussion. For instance, the annotator could use it to convince an opponent about the given claim. For example, both tweets "Petition to stop calling people who don't wear masks "anti-maskers". Instead, let's just call them terrorists. #coronavirus" and "Masks help prevent the spread of the disease. Please, #WEARAMASK" are in favor of the claim, yet only the second tweet contains the argument. The tweets were manually annotated for binary classification, and participants of this subtask were required to submit whether each tweet has a premise (1) or not (0).

We split an existing corpus of 4,269 COVID-related tweets (Glandt et al., 2021) into a training set of 3,669 tweets and a validation set of 600 tweets. This corpus includes annotations for Task 2a. We added a new annotation layer for Task 2b to this corpus. In order to create the test set, we collected new tweets using 33 keywords such as #OpenSchools, #LockdownNow, #NoMasks. We removed the hashtags from the tweets to exclude obvious signals (e.g., #SayNoToMasks can be seen as the "against" hashtag). The test set for Task 2a and all three sets for Task 2b were manually annotated. Three *Yandex.Toloka*[1] annotators' crowd-sourced labels were aggregated into a single label (Dawid and Skene, 1979). Table 1 shows statistics of the experimental datasets. More details on the datasets are presented in (Davydova and Tutubalina, 2022). We used the F_1-score as the main evaluation metric in both sub-tasks: $F_1 = \frac{1}{n}\sum_{c \in C} F1_{rel_c}$, where C ={"face masks", "stay at home orders", "school closures"}, n is the size of C, $F1_{rel_c}$ is macro F_1-score averaged over two classes for each task (in-favor & against classes for Task 2a; 0 & 1 classes for Task 2b). The Task 2 is hosted on Codalab at https://codalab.lisn.upsaclay.fr/competitions/5067.

[1] https://toloka.yandex.ru/

2.3 Task 3: Classification of changes in medication treatments in tweets and WebMD reviews (in English)

In Task 3, we challenged the participants to design binary classifiers to detect posts where social media users self-declare changing their medication treatments, regardless of being advised by a health care professional to do so. Such changes are, for example, not filling a prescription, stopping a treatment, changing a dosage, forgetting to take the drugs, etc. This is an important task since it is the first step toward detecting patients who may be non-adherent to their treatments and it would allow us to further understand their reasons when they are expressed on social media (Weissenbacher et al., 2021). We released two corpora for the task, 9,830 tweets from Twitter and 12,972 drug reviews from WebMD, a website which provides an opportunity for users to comment anonymously on their personal experience with a drug in a free text form. Two annotators labeled the posts with "1" when the posts state a change in medication regimen, the positive examples, "0" otherwise, the negative examples. The Inter-annotator agreements were moderate on both corpora with 0.65 and 0.74 Cohen Kappa scores on the corpus of tweets and reviews, respectively. With 7,222 positive examples and 5,750 negative examples, the corpus of WebMD is naturally balanced with an Imbalance Ratio of 0.80. The corpus of tweets is more challenging for training classifiers with supervision, which is the default approach to solve the task. The corpus of tweets has a strong imbalance with only 864 positive examples for 8,966 negative examples, that is a 10.38 Imbalance Ratio. We split randomly each corpus into a training, a validation, and a test set with 90%/10%/10% of the total reviews available and 60%/16%/24% of the total tweets available in each set. During the practice period of the competition, we provided the participants the training and validation sets. During the evaluation period, all participants had 5 days to compute their predictions on the test set and submit them on Codalab for evaluation. We added 11,835 reviews and 13,000 tweets as decoys in the test sets to prevent manual correction of the predictions. We evaluated participants' systems with the Precision, Recall and F_1-score for the positive class, that is tweets or reviews mentioning a change in medication treatments. The Task 3 is hosted on Codalab at https://codalab.lisn.upsaclay.fr/competitions/2138.

2.4 Task 4: Classification of tweets self-reporting exact age (in English)

A limitation of using Twitter data for research applications is that users may not be representative of the general population. Therefore, advancing the utility of Twitter data for research applications requires methods for automatically detecting demographic information, including users' age. Automatically identifying the exact age of Twitter users, rather than their age groups, would enable the large-scale use of Twitter data for applications that do not align with the predefined age groupings of extant models, including health applications such as identifying specific age-related risk factors for observational studies (Golder et al., 2019), or selecting age-based study populations (Davies et al., 2022). As a first step, this binary classification task involves automatically distinguishing tweets that self-report the user's exact age ("age" tweets) from those that do not ("no age" tweets). The training set contains 8800 annotated tweets: 2834 (32%) "age" tweets and 5966 (68%) "no age" tweets. The validation set contains 2200 annotated tweets: 709 (32%) "age" tweets and 1491 (68%) "no age" tweets. The test set also contains 2200 annotated tweets: 768 (35%) "age" tweets and 1432 (65%) "no age" tweets. Inter-annotator agreement (Fleiss' kappa), based on 1000 tweets annotated by five annotators, was 0.80. The Task 4 is hosted on Codalab at https://codalab.lisn.upsaclay.fr/competitions/3566.

2.5 Task 5: Classification of tweets containing self-reported COVID-19 symptoms (in Spanish)

The purpose of this task is to bridge the gap in NLP and social media for COVID-19 research performed in languages other than English. While there has been an increased amount of non-English datasets and tasks using social media proposed in the last couple of years, there is still a need for different applications on pressing topics. This shared task is similar to the 2021 #SMM4H shared task 6 (Magge et al., 2021), which involved identifying personal mentions of COVID-19 symptoms tweets. The annotated set of tweets for this task is a set of manually curated Spanish-native language tweets. The task is a three-way classification problem, requiring participants to distinguish personal symptom mentions (self-reports) from other mentions such as symptoms reported by others (non-personal reports) and references to external sources (literature/news mentions). We provided a training dataset consisting of 1,654 tweets labeled as self-reports, 2,413 tweets labeled as non-personal reports, and 5,985 labeled as literature/news mentions. The test set consisted of 6,851 tweets, 1,096 self-reports, 1,644 non-personal reports, and 4,111 literature/news mentions. The complete annotated dataset (train, validation, testing) has a Cohen Kappa score inter annotator agreement of 0.85. The systems submitted for this task where evaluated using micro-average F1-score. The Task 5 is hosted on Codalab at https://codalab.lisn.upsaclay.fr/competitions/3535.

2.6 Task 6: Classification of tweets which indicate self-reported COVID-19 vaccination status (in English)

With the widespread roll-out of COVID-19 vaccines, vaccine surveillance became a very pressing research issue. While some vaccinated people report adverse events via their healthcare providers to systems like Vaccine Adverse Event Reporting System (VAERS), or are found documented in their electronic health record (EHR), a more robust and convenient method could be devised using self-reports from social media. In this task we provided an annotated dataset of Tweets with users personally reporting vaccination status or discussing vaccination status but not revealing their own, extracted from Banda et al. (Banda et al., 2021). This task is challenging since users often discuss vaccination status of others or from news reports in similar ways and at a higher rate than they discuss their own (1 to 8 on average). The dataset presents as the positive class, unambiguous tweets of users clearly stating that they have been vaccinated (vaccination confirmation). All other tweets are of users discussing vaccination status. This task involved the identification of self-reported COVID-19 vaccination status in English tweets (vaccine related chatter). This task is posed as a two-way classification task, where the systems submitted were evaluated on precision, recall and F1-score. The class imbalance in this dataset is roughly 1 to 8, meaning that for training we provided 1,496 tweets of vaccination confirmation, and 12,197 of vaccine chatter tweets. The test set is comprised of 652 tweets labeled as self-reports and 5,271 tweets of vaccine chatter. The complete annotated dataset (train, validation, testing) has a Cohen Kappa score inter annotator agreement of 0.82. The Task 6 is hosted on Codalab at https://codalab.lisn.upsaclay.fr/competitions/3536.

2.7 Task 7: Classification of self-reported intimate partner violence on Twitter (in English)

Intimate partner violence (IPV), which refers to abuse or aggression that occurs in a romantic relationship, is a serious health problem that can have a lifelong impact on health and well-being (Smith et al., 2018). Social media platforms are often used by IPV victims to share experiences and seek for help (Westbrook, 2015; Cravens et al., 2015; McCauley et al., 2018; Chu et al., 2021; Al-Garadi et al., 2022). To potentially provide timely intervention and support to victims, we have the need for an effective automatic classifier to detect self-reports of IPV on social media platforms. Task 7 is a binary classification task that involves identifying the IPV self-report posts on Twitter. This task presents two specific challenges. First, the annotated data is significantly imbalanced where only around 11% of the tweets are identified as self-reports of IPV. Second, the negative tweets include non-IPV domestic violence and non-self-reported IPV, which can be very difficult to distinguish from self-reported IPV for an automatic system. The data (Al-Garadi et al., 2022) include a total of 6,348 annotated posts from Twitter, of which 4,523 were provided for training, 534 provided for validation, and 1,291 provided for testing. IAA was found to be 0.86 (Cohen's kappa) among 1,834 double-annotated tweets. Systems were evaluated and ranked based on the F_1-score of the positive class (self-reported IPV). The Task 7 is hosted on Codalab at https://codalab.lisn.upsaclay.fr/competitions/1535.

2.8 Task 8: Classification of self-reported chronic stress on Twitter (in English)

Chronic stress is defined as the *physiological or psychological response to a prolonged internal or external stressful event* (VandenBos, 2007), which can lead to poor mental health, including depression and anxiety, and can also take a toll on the body, resulting in the dysfunctions of cardiovascular, metabolic, endocrine, and immuno-inflammatory systems. Traditional methods for assessing stress, including interviews, questionnaires/surveys, etc., have limitations associated with accurately measuring stress at the population level (Epel et al., 2018). Therefore, there is a need to identify new sources of data and new methods for assessing chronic stress. One potential source of information for analyzing chronic stress related information is social media such as Twitter. The first step to do so is to accurately detect the tweets that report personal experiences of chronic stress. Task 8 is a binary classification that involves automatically identifying tweets that are self-disclosures of chronic stress from those that are not.

For Task 8, we released a corpus of tweets where about 37% of the tweets are positive (self-disclosures; P) and 63% are negative (non-self-disclosures; N). We split the corpus in three set, the training set which contains 2,936 tweets, the validation set, 420 tweets, and the test set, 839 tweets. The pairwise inter-annotator agreement among 695 double-annotated tweets was κ=0.83 (Cohen's kappa (Cohen, 1968)), which can be interpreted as a substantial agreement. Further details about the data and the annotation process are provided in Yang et al. (2022b). Classifiers were evaluated based on the F_1-scores for the "positive" class (i.e., tweets that are self-disclosure of chronic stress). The Task 8 is hosted on Codalab at https://codalab.lisn.upsaclay.fr/competitions/1542.

2.9 Task 9: Classification of Reddit posts self-reporting exact age (in English)

Pharmaceutical companies, with the encouragement of regulatory agencies (Donegan et al., 2019; U.S. Food and Drug Administration, 2020), have started using social media listening (SML) to integrate the patient perspective in the clinical development process to ensure relevant treatments and outcomes. Traditionally, SML in the pharmaceutical setting has been done through manual, qualitative methods. However, it has been shown that quantitative SML (QSML) can enhance the value of social media data and enable a patient-centric approach to understanding disease burden and influence drug discovery decisions at all stages (Schmidt et al., 2022).

The detection of self-reported demographic information on social media can help in assessing the demographic characteristics (e.g. age, gender, ethnicity, medical history) of patients on social media in comparison to patients in target clinical populations. Task 9 is a binary classification that aims to distinguish automatically posts in Reddit forum where users that self-report their exact age at the time of posting from those where they do not. The dataset is disease-specific and consists of posts collected via a series of keywords associated with dry eye disease. The training set contains 9,000 annotated posts mentioning numbers that were randomly selected: 2,921 (32.5%) with self-reported ages (annotated as "1") and 6,079 (67.5%) with no self-reported ages (annotated as "0"). The test set contains 2,000 annotated posts: 629 (31.5%) with self-reported ages (annotated as "1") and 1,371 (68.5%) with no self-reported ages (annotated as "0"). Inter-annotator agreement (Cohen's kappa) was 0.939. Systems were evaluated based on the F1-score for the target class (self-reported age mentioned). The Task 9 is hosted on Codalab at `https://codalab.lisn.upsaclay.fr/competitions/3646`.

2.10 Task 10: SocialDisNER - Detection of disease mentions in tweets (in Spanish)

Since diseases is an important category of named entities to help recognizing health-related content in social media, the community has invested time and effort to collect and annotate large corpora to train Named-Entity Recognition systems to perform the task automatically. However, most corpora are written in English, for example (Scepanovic et al., 2020), CADED (Karimi et al., 2015), Micromed (Jimeno-Yepes et al., 2015), or TwiMed (Alvaro et al., 2017). Fewer corpora written in Spanish, like SpanishADR (Segura-Bedmar et al., 2014) or ProfNER (Miranda-Escalada et al., 2021), are available. Task 10, hereafter called SocialDisNER, is a first attempt to fill the gap. The goal of SocialDisNER is the automatic recognition of disease mentions in tweets. We used the LINKAGE methodology (Gasco et al., 2022) to select a set of disease-related tweets, written in Spanish and with first-hand experience from patients and their relatives. The corpus also contains relevant health information tweets written by medical professionals. The corpus consists of 9,500 tweets, with 5,000 tweets used for training, 2,500 for development and 2,000 for evaluation. The corpus was annotated by a medical professional following an adaptation of the DisTEMIST annotation guidelines, which were tested with several rounds of inter-annotation agreement (IAA) to achieved a final human-IAA of 0.823 (Gasco et al., 2022). The primary evaluation metric for the task was the micro-averaged F1-score. Participants were also compared to a baseline system that was computed using a Levenshtein lexical lookup approach with a sliding window of variable length. In addition to the Gold Standard of 9,500 tweets, we also provided the participants with a Silver Standard. It is a large-scale corpus of 80,000 tweets where we automatically annotated the mentions of diseases, symptoms, procedures, drugs, species and professions, among others. Annotation was carried out through NER systems trained on clinical data in Spanish and post-processing to eliminate false negatives such as URLs or twitter mentions. In order to encourage the use of the large-scale corpus and foster the use of mined content, the co-mention matrices of the Large scale corpus were calculated and shared. These matrices represent the co-occurrences of disease mentions to each other, as well as the co-occurrences of diseases with other terms such as symptoms or professions. An example of these matrices can be seen in the co-mention network in Figure 1. The Task 10 is hosted on Codalab at `https://codalab.lisn.upsaclay.fr/competitions/3531`.

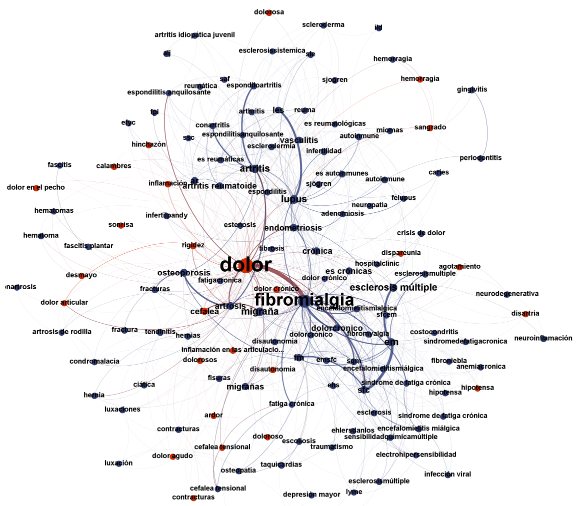

Figure 1: Simplified SocialDisNER co-mention network between disease and symptoms. Network filtered for fibromyalgia symptoms.

3 Results

3.1 Task 1: Classification, detection and normalization of Adverse Events (AE) mentions in tweets (in English)

A total of 34 teams participated in at least one of the subtasks of Task 1 making 80 valid submissions. 19 out of the 34 teams submitted system descriptions. Tables 2, 3, 4 presents F_1 scores for the participating teams' best submissions along with brief descriptions of the architecture choices made by the teams. We encourage readers to refer to the original papers for detailed descriptions of the systems. Similar to last years' submissions, all teams for Task 1a used transformer models, while for Tasks 1b and 1c, about 70% of the submissions used transformer models. RoBERTa and BERTweet (based on RoBERTa) were the most popular choice of models and model ensembles were used in many of the top submissions. Variations of off-the-shelf models were used in the tasks for addressing the class imbalance problem and the large label space problem presented in the normalization task. Some of these variations included data augmentation techniques, multi-corpus training, optimizer adaptations and adversarial training.

3.2 Task 2: Classification of stance and premise in tweets about health mandates related to COVID-19 (in English)

Table 5 presents F_1 scores for each of the 15 team's best-performing system. Most teams used COVID-related BERT models with additional techniques, such as regularized dropout (R-drop), to alleviate the unbalanced label distribution and overfitting. Two teams (# 22, 28) tried to combine the data to train a model that perform simultaneously both tasks. Leading teams on both tasks tried to aggregate claim and tweet texts: the leading team #7 with 0.64 F_1 in Task 2a appended the claim text to the end of the tweet, while the second-best team #12 in Task 2a with highest F_1 (0.7) in Task 2b proposed a new pre-training task constructed by the tweets and claims similarly to next sentence prediction.

3.3 Task 3: Classification of changes in medication treatments in tweets and WebMD reviews (in English)

We observed a good interest in Task 3 with 43 teams registered. Among these teams, 7 teams submitted their predictions to Codalab. In Table 6, we summarized their performance and compare them with our baseline systems described in (Weissenbacher et al., 2021). The main focus of the participants during the competition was the generalizability of their classifiers. These past years, the community released large pre-trained embeddings, such as glove, fasttext, or transformers, along with convenient interfaces to integrate them into neural networks. These interfaces propose default architectures for these neural networks that can be programmatically customized by advanced users. Such networks provide new opportunities since they can be easily trained with supervision to solve any classification tasks at hand by fine-tuning their models on small - or larger - annotated corpora and still achieving competitive results. All but one participants took advantages of these neural networks and evaluate their abilities to solve multiple tasks of the #SMM4H 2022 competition, among which Task 3. We note that few teams adapted their training process to improve performances on imbalance corpora, which was the key of success for sub-task 3a. This corpus of tweets remains challenging with the best scores plateauing around 0.65 F_1-score compare to the balanced corpus of WebMD reviews were transformer-based approaches achieved high F_1 scores, scores above 0.80.

3.4 Task 4: Classification of tweets self-reporting exact age (in English)

Table 7 presents the precision, recall, and F_1-score for each team's best-performing system for Task 4. The four top-performing systems achieved F_1-scores within less than 0.01 points of one another using BERTweet or RoBERTa pretrained transformer models. Among these four systems, using a model pretrained on tweets did not seem to improve performance for this task. These four teams marginally outperformed a RoBERTa-Large baseline classifier (Klein et al., 2022) by using techniques such as pseudo-labeling, ensemble learning, multi-corpus training, adversarial attacks, and child-tuning. All seven systems that used BERTweet or RoBERTa models, including the baseline classifier, outperformed systems that used BERT or BioBERT models.

3.5 Task 5: Classification of tweets containing self-reported COVID-19 symptoms (in Spanish)

With 26 participants registered in CodaLab and seven final team submissions highlighted on Table 8, this task presented an interesting challenge to the teams. Most teams, with distinct approaches ranging from ensemble models to tuned BERT-based model, performed well and only 3 percentage points separated the best from the last ranking teams. The baseline model, which featured a hefty augmented training set using weak supervision (Tekumalla and Banda, 2021), still man-

Team	F_1	P	R	System Summary
25	0.698	0.839	0.598	BERTweet-large with data augmentation
23	0.693	0.772	0.629	Majority ensemble of 10 RoBERTa-large models
45	0.689	0.790	0.611	DeBERTa-v3-large model with adversarial training
-	0.671	0.799	0.578	-
-	0.669	0.791	0.579	-
14	0.662	0.785	0.573	RoBERTa and BERTweet with exponential moving average
-	0.662	0.790	0.570	-
46	0.662	0.765	0.584	Ensemble of BERTweet, DeBERTa and BioBERT with data augmentation
-	0.660	0.787	0.569	-
37	0.655	0.688	0.625	Ensemble modeling from T5 gpt2-large templates with over/under sampling
41	0.652	0.737	0.585	RoBERTa model pretrained on in-domain tweets
11	0.642	0.554	0.765	Glove embeddings and DeepADEMiner (RoBERTa) classifier
-	0.638	0.807	0.528	-
10	0.637	0.787	0.536	Fine-tuned RoBERTa-base with Adversarial Training
27	0.610	0.606	0.614	Ensemble of finetuned BERTweet-large, RoBERTa-large, CT-BERT
40	0.601	0.705	0.524	RoBERTa with Fast Gradient Method and Project Gradient Descent
38	0.567	0.674	0.489	RoBERTa with adaptive learning and mixut
-	0.537	0.724	0.427	-
18	0.491	0.384	0.681	RoBERTa with data augmentation and downsampling techniques
28	0.472	0.607	0.386	BERT fine-tuned on medically relevant data
-	0.470	0.659	0.365	-
8	0.433	0.614	0.334	Template-augmented training using BERTweet model
17	0.413	0.677	0.297	BERT, RoBERTa and ERNIE 2.0 with voting ensembler
-	0.396	0.593	0.297	-
-	0.333	0.398	0.287	-
-	0.326	0.603	0.223	-
1	0.299	0.235	0.409	Ensemble of BERT, BioBERT, XLNet, RoBERTa
-	0.224	0.485	0.145	-
36	0.077	0.041	0.547	RoBERTa and BERTweet with label-distribution aware margin loss

Table 2: Evaluation results for Task 1a: Classification of ADE mentions in English tweets. Metrics show F_1-scores (F_1), precision (P), and recall (R) for the *ADE* class.

Team	F_1	P	R	System Summary
45	0.651	0.684	0.621	W2NER model with character and location features
-	0.642	0.688	0.601	-
-	0.640	0.686	0.600	-
-	0.639	0.683	0.599	-
-	0.637	0.681	0.599	-
11	0.624	0.569	0.691	DeepADEMiner(RoBERTa-large) and Flair embeddings
23	0.568	0.671	0.492	Majority ensemble of 5 BERT-large models
37	0.519	0.644	0.434	Ensemble modeling from T5 gpt2-large templates with over/under sampling
25	0.484	0.828	0.341	BERTweet-large with positive examples from WEBRADR dataset
38	0.435	0.562	0.354	Question Answering methodology using RoBERTa-base fine-tuned on SQuAD2.0 dataset
28	0.404	0.489	0.344	Fine-tuned BERT developed for SMM4H 2019
1	0.220	0.178	0.288	Fine-tuned RoBERTa+CRF model
-	0.164	0.096	0.576	-
-	0.132	0.074	0.651	-

Table 3: Evaluation results for Task 1b: Extraction of ADE mentions in English tweets. Metrics show overlapping F_1-scores (F_1), precision (P), and recall (R) for the *ADE* spans.

Team	F_1	P	R	System Summary
11	0.387	0.350	0.432	DeepADEMiner(bert-based-uncased) and mcn-en-smm4h model (BioBERT)
45	0.367	0.405	0.336	Ranked average pooling of DeBERTa model word vectors
28	0.243	0.294	0.207	GPT-2 model trained to predict MedDRA term from ADE span
38	0.172	0.232	0.137	Fuzzy matching with Levenshtein distance
1	0.116	0.094	0.152	Ranked similarity metrics of extracted spans and MedDRA dicitionary
23	0.070	0.087	0.058	Stop-word removal and simple string matching in MedDRA
-	0.013	0.019	0.009	-
-	0.011	0.017	0.008	-
-	0.010	0.015	0.007	-
-	0.008	0.013	0.006	-
-	0.007	0.011	0.005	

Table 4: Evaluation results for Task 1c: Extraction and normalization of ADE mentions in English tweets. Metrics show overlapping F_1-scores (F_1), precision (P), and recall (R) for the *ADE* spans with exact matches for MedDRA preferred term IDs.

Team	F1 stance (Task 2a)	F1 premise (Task 2b)	System Summary
7	**0.64**	0.66	CovidTwitterBERT (for 2a) / BART-base (for 2b) + cross-entropy and contrastive losses + features
12	0.63	**0.70**	COVID-Twitter-BERT-v2 / RoBERTa-large + Tweet Claim Matching (TCM) pre-training
22	0.62	0.62	CT-BERT V2 + related data, sentiment classification along with multi-task learning
42	0.62	0.63	Dual-view attention neural network + COVID-Twitter-BERT-v2
14	0.59	**0.70**	BERTweet + regularized dropout (R-drop), exponential moving average (EMA), and focal loss
23	0.58	**0.70**	RoBERTa-large + majority ensemble (2a), BERT-large + weighted average ensemble (2b)
46	0.55	0.64	BioBERT + R-drop, poly loss and focal loss, pseudo labels
28	0.53	0.65	6-way joint classification + ensemble of RoBERTa and BERT
17	0.50	0.62	A voting-ensemble model that comprises fine-tuned BERT, RoBERTa, and ERNIE 2.0 / fine-tuned single models
31	0.48	0.64	A voting classifier to ensemble the predictions of BERT, RoBERTa, PubMedBERT, SciBERT and SPECTER
19	0.43	0.63	bioBERT-base
-	0.32	-	-
9	0.23	0.23	An ensemble of fine-tuned GAN-BERT models
6	0.08	0.00	RoBERTa-large/BERTweet-large
29	-	0.67	SqueezeBERT

Table 5: Evaluation results for Task 2: Classification of stance and premise in tweets about health mandates related to COVID-19. The best scores for each task are in bold and second best scores are underlined.

aged to rank the highest with a micro F_1-score of 0.90. Team 14 and their pre-trained BERT-base model with R-drop, exponential moving average, and pseudo-labeling was the highest ranking team with a micro F_1 score of 0.86.

3.6 Task 6: Classification of tweets which indicate self-reported COVID-19 vaccination status (in English)

The task of classifying self-reported COVID-19 vaccination status, in English, attracted 44 participants in the competition, with eight teams submitting their predictions on the unseen test set on Codalab. Table 9 shows that team 28 managed to match the baseline system as the best performing systems, with an improvement in precision and a

Team	F_1	P	R	System Summary
Task 3a				
14	**0.66**	0.68	**0.63**	Ensemble of BERTweet classifiers trained with 5-fold cross validation + Cost sensitive learning and data augmentation
46	0.64	**0.68**	0.60	BERTweet + data augmentation
33	0.61	0.66	0.57	RoBERTa embeddings input for a neural network with a stacked LSTM and linear layer unified branches
17	0.59	0.62	0.56	Bio-RoBERTa
baseline	0.50	0.47	0.53	CNN trained with transfer and active learning
19	0.45	0.53	0.39	BioBERT
11	0.19	0.55	0.12	Glove embeddings pretrained on tweets input for Bi-LSTM + Cost sensitive learning
41	0.06	0.03	0.34	RoBERTa embeddings pretrained on tweets
Task 3b				
baseline	**0.87**	**0.87**	0.88	BERT-based
33	0.86	0.85	0.88	RoBERTa embeddings input for a neural network with a stacked LSTM and linear layer unified branches
46	0.86	0.86	0.85	BERTweet + data augmentation
19	0.83	0.84	0.82	BioBERT
41	0.72	0.57	**1.0**	RoBERTa embeddings pretrained on tweets

Table 6: System summaries and scores (F_1), precision (P), and recall (R)) for Task 3: Classification of changes in medication treatments in tweets and WebMD reviews - in English

Team	F_1	P	R	System Summary
46	**0.92**	**0.93**	**0.91**	BERTweet, pseduo-labeling, R-Drop, PolyLoss
10	0.92	0.93	0.90	RoBERTa-Base, sub-corpus ensemble, adversarial training, child-tuning
14	0.91	0.92	0.91	Ensemble of BERTweet classifiers trained with 5-fold cross validation, pseudo-labeling, R-Drop, EMA
6	0.91	0.92	0.90	RoBERTa-Large, additional training data from Task 9
Baseline	0.91	0.93	0.88	RoBERTa-Large
18	0.89	0.90	0.89	RoBERTa
35	0.85	0.80	0.89	RoBERTa-Large
15	0.82	0.77	0.87	BERT-Base-Uncased
19	0.81	0.82	0.80	BioBERT-Base-Cased
-	0.73	0.67	0.79	-

Table 7: System summaries and F_1-scores (F_1), precision (P), and recall (R) for Task 4: Classification of tweets self-reporting exact age (in English).

Team	F_1	P	R	System Summary
baseline	**0.90**	**0.90**	**0.90**	CT-BERT + data augmentation (labeled using weak supervision)
14	0.86	0.86	0.86	Custom pre-training BERT-based model + R-drop, exponential moving average, and pseudo-labeling
26	0.85	0.85	0.85	XLM-RoBERTa-base + post-processing step
46	0.85	0.85	0.85	BioBERT + R-drop + focal loss
13	0.85	0.85	0.85	Majority ensemble of: BERT BASE-multilingual, BETO, XLM-RoBERTa
5	0.84	0.84	0.84	Fine-tuned RoBERTuito + data pre-processing
28	0.84	0.84	0.84	Ensemble of two differently configured BERT + one RoBERTa models
17	0.83	0.83	0.83	XLM-R

Table 8: Evaluation results for Task 5: Classification of tweets containing self-reported COVID-19 symptoms (in Spanish). Metrics shows F_1-scores (F_1), precision (P), and recall (R) for the self-reports class. The table shows the best submission of each team and their system.

drop in recall using an ensemble model. Similar to other classification shared tasks, most systems used BERT-based models with variants like CT-BERT and BioBERT being popular. The two approaches, baseline and team 27 with augmented data performed better (on average) than the systems that did not use additional data, so this is an interesting result from the presented approaches.

3.7 Task 7: Classification of self-reported intimate partner violence on Twitter (in English)

Table 10 presents F_1-scores, precision, and recall for the Intimate Partner Violence (IPV) self-report class for the participating teams. The median F_1-score, precision, and recall of all the submissions are 0.763, 0.79, and 0.716, respectively. All of the participants used pre-trained transformer-based models, 4 out of 7 teams used multiple BERT variants, 4 teams used RoBERTa, and 3 teams used ensemble techniques. The leading team achieved F_1-score of 0.851 which is 0.8 higher than the baseline system. The leading team also achieved the best recall, which is 0.4 higher than the second highest recall score. The leading team used RoBERTa-Large to encode tweet text and make a binary prediction according to the corresponding pooling vector. The leading team trained on 5 RoBERTa models and applied a weighted ensemble strategy. The leading team achieved lower precision but higher recall than the second-placed team, which also used RoBERTa and domain-adaptive pre-training. The leading team achieved higher precision and recall than those of the third-placed team, which applied an averaging ensemble on different transformer-based models. Team 3 and 10 achieved significantly higher F_1-scores compared to other teams, which might be attributed to the efficient weighted ensemble strategy and domain adaptive pre-training.

3.8 Task 8: Classification of self-reported chronic stress on Twitter (in English)

Table 11 presents the F_1-scores, precision, and recall for the positive class, for each of the 11 team's best-performing system for Task 8. The best performance achieved in task 8 was an F_1-score of 0.792, which is comparable to the benchmark reported in the literature on the same corpus (Yang et al., 2022b). The median F_1-score, precision and recall of all submissions are 0.75, 0.72 and 0.76, respectively. Only 7 of 11 teams submitted their system descriptions, among which 6 of the 7 teams used RoBERTa. 3 teams used ensemble systems built from multiple pre-trained transformer-based models, among them, 2 teams included BERT and 2 teams included BERTweet. The leading team preprocessed the texts by lowercasing, deleting URLs, replacing emojis with their text strings; and used 5-fold CV during the training phase. The leading team used three pre-trained BERT-based models including BERT, RoBERTa and BERTweet then fine-tuned them using task 8's datasets. To further improve the models' performance, the leading team also adopted pseudo-labeling in the post-processing phase. The leading team's results showed that BERTweet outperformed BERT and RoBERTa for this task. Both top teams included BERTweet in their systems, indicating the benefit of pre-training the embeddings on social media based corpora for this task.

Team	F_1	P	R	System Summary
baseline	**0.83**	**0.9**	**0.77**	CT-BERT + data augmentation (labeled using weak supervision)
28	0.83	0.93	0.75	Ensemble of three differently configured BERT models.
27	0.82	0.86	0.78	CT-BERT fine-tuned from scratch + Data augmentation (manual and automatic)
46	0.81	0.90	0.74	BioBERT + R-drop
14	0.80	0.90	0.71	Custom pre-training BERT-based model + R-drop, exponential moving average, and pseudo-labeling
10	0.78	0.91	0.68	Continued pre-trained RoBERTa$_basemodel$
11	0.69	0.87	0.87	LSTM + GLOVE Twitter Embeddings
17	0.68	0.76	0.87	BERT
47	0.66	0.77	0.93	Voting ensemble of: fine-tuned BERT, RoBERTa, and XLNet models

Table 9: Evaluation results for Task 6: Classification of tweets which indicate self-reported COVID-19 vaccination status (in English). Metrics shows F_1-scores (F_1), precision (P), and recall (R) for the self-reports class. The table shows the best submission of each team and their system.

Team	F_1	P	R	System Summary
3	0.851	0.86	0.841	RoBERTa-Large with weighted ensemble
10	0.833	0.875	0.795	RoBERTa with domain adaptive pre-training
14	0.795	0.795	0.795	BERT, RoBERTa, and BERTweet with averaging ensemble
29	0.791	0.903	0.705	Domain specific BERT variants with different loss functions
Baseline	0.756	0.823	0.699	RoBERTa
46	0.734	0.784	0.689	Six BERT variants including RoBERTa with voting ensemble
19	0.707	0.763	0.659	BioBERT
36	0.625	0.549	0.727	RoBERTa and BERTweet with different loss functions

Table 10: Evaluation results for Task 7: Classification of self-reported intimate partner violence on Twitter (in English). Metrics show F_1-scores (F_1), precision (P), and recall (R) for the self-reports class.

Team	F_1	P	R	System Summary
14	**0.792**	0.734	0.859	BERT, RoBERTa, and BERTweet with averaging ensemble
46	0.783	**0.797**	0.769	BioBERT, pubmedBERT, DeBERTa and BERTweet with different loss functions
10	0.781	0.739	0.827	RoBERTa with domain adaptive pre-training
-	0.773	0.731	0.821	-
-	0.764	0.793	0.737	-
19	0.764	0.677	0.878	RoBERTa
44	0.750	0.718	0.785	RoBERTa with BiLSTM
21	0.730	0.703	0.760	BERT, RoBERTa, ALBERT, XLNet and ELECTRA transformers
-	0.719	0.743	0.696	-
-	0.643	0.599	0.692	-
41	0.542	0.372	**1.000**	RoBERTa

Table 11: Evaluation results for Task 8: Classification of self-reported chronic stress on Twitter. Metrics show F_1-scores (F_1), precision (P), and recall (R) for the self-disclosure class.

3.9 Task 9: Classification of Reddit posts self-reporting exact age (in English)

Table 12 presents the precision, recall, and F_1-score for each team's best-performing system for Task 9. The median F1-score, precision, and recall of all the submissions were 0.891, 0.896 and 0.919, respectively. All participating teams had a common approach which was to create a generic framework to fine-tune pre-trained transformer-base models, with which they participated in multiple classifications tasks. Two teams (6 and 35), nonetheless, focused their attention only on tasks 4 and 9, which were highly similar. These 2 teams obtained contrasting results, with one being the winner of the competition and the other being second to last. The best submissions of each team ended up using RoBERTa (3 teams), BERTweet (2 teams), BioBERT (1 team) and BERT-Large (1 team). Furthermore, 2 teams used ensemble models and 4 teams used a combination of additional strategies to improve performance (R-drop, FocalLoss, pseudo-labeling, oversampling, among others).

The best performance was obtained by team 6 with an F_1-score of 0.956, including the best recall, 0.963. The winner team used RoBERTa-Large with an augmented training corpus that included the training data from Task 4. This team was the only team that augmented the training dataset.

Finally, 2 teams additionally performed error analysis on the validation set. One of them found sufficient mislabeled posts, both as false positives and false negatives, that suggest that the dataset could benefit from a revision. The other team found that, unsurprisingly, their model puts special focus on numbers which might lead to incorrect predictions as it fails to understand complex semantics where the age is indirectly reported.

3.10 Task 10: SocialDisNER - Detection of disease mentions in tweets (in Spanish)

SocialDisNER has achieved good participation results with a total of 47 registered teams and 17 participating teams. Table 13 shows the ranking of the teams that uploaded predictions to the task according to their micro-average F1-score. There were 11 teams that achieved better performance than the baseline system. The top-performing team (Fu et al., 2022) obtained an F_1-score of 0.891 by developing a Unified Named Entity Recognition system through domain-adaptive pre-training using a domain-general Spanish BERT model (Canete et al., 2020) fine-tuned by adversarial training, child adjustment and model fusion.

Most participants used systems built from pre-trained transformer-based models and obtained the final predictions using token classification layers, CRFs or ensembles. A significant number of the participants (6 out of 17) have used or tested some of the language models trained with biomedical texts in Spanish published in the last few months (Carrino et al., 2022; Chizhikova et al.; Lange et al., 2021). Two teams opted to use general domain models in Spanish (Canete et al., 2020; Gutiérrez Fandiño et al., 2022) tuned using domain adaptation or weak training. A total of six teams chose to use multilingual models such as XLM-RoBERTa or multilingualBERT (Devlin et al., 2018; Conneau et al., 2019), one of them finishing in the top 3, showing how a good fine-tuning and processing strategy can deliver reliable results. It is worth highlighting that several teams used models that were pretrained on Spanish tweets (Huertas-Tato et al., 2022; Barbieri et al., 2022), although, they all achieved moderate performances. A possible explanation for these results may be that, to efficiently extract biomedical entities in Tweets, the models needs to transfer the linguistic knowledge learned during their pre-training on biomedical content, a knowledge which seems missing in general domain content like general tweets.

Six teams used the SocialDisNER large-scale corpus to solve the task. The top-performing team (Fu et al., 2022) used it to apply continual pre-training on the Spanish generalist language model to achieve better adaptation to the task domain. SINAI team (Chizhikova et al., 2022) used it to carry out a weak-supervision approach when training their system.

Most teams used the DisteMIST disease gazetteer for preprocessing and postprocessing disease mentions contained within special tweet tokens such as hashtags. In addition, they also used other resources such as CodiESP data (Miranda-Escalada et al., 2020) or biomedical resources such as Snomed-CT, UMLS or ICD-10 terminologies.

4 Conclusion

This paper presented an overview of the #SMM4H 2022 shared tasks. With ten tasks proposed this year, the interest and the participation in the #SMM4H shared tasks continues to grow. All best

Team	F_1	P	R	System Summary
6	**0.956**	0.948	**0.963**	RoBERTa-Large augemented with Task 4 training data
46	0.938	**0.957**	0.919	BERTweet and DeBERTa with R-drop, FocalLoss, PolyLoss and pseudo-labeling
17	0.918	0.896	0.941	Fine-tuned RoBERTa
14	0.891	0.893	0.889	BERTweet with averaging ensemble wirth R-drop, Exponential Moving Average, FocalLoss and pseudo-labeling
10	0.885	0.909	0.862	Continue RoBERTa-Base with averaging sub-corpus ensemble, FGM adversarial training, child-tuning and oversampling
35	0.865	0.797	0.946	BERT-Large with Synthetic Minority Oversampling Technique
19	0.856	0.862	0.851	BioBERT

Table 12: Evaluation results for Task 9: Classification of Reddit posts self-reporting exact age. Metrics show F_1-scores (F_1), precision (P), and recall (R) for the self-reported age class. The table shows the best submission of each team and their system.

Team	P	R	F1	System summary
10	**0.906**	<u>0.876</u>	**0.891**	Spanish BERT with domain adaptive pre-training + adversarial training + child tuning + model fusion
43	0.868	0.875	<u>0.871</u>	RoBERTa-base trained on biomedical corpus in Spanish + post-processing
39	0.851	**0.888**	0.869	Pre-processing and gazetteer lookup + XLM RoBERTa-Large
30	<u>0.882</u>	0.843	0.862	RoBERTa-base trained on biomedical corpus in Spanish with contextualized embeddings at document level.
34	0.842	0.860	0.851	Multilingual BERT + post-processing
26	0.828	0.845	0.836	XLM-RoBERTa-base + rule-based system
2	0.868	0.779	0.821	XLM-RoBERTa-base + lateral inhibitory layer
20	0.809	0.798	0.803	RoBERTa-base trained on biomedical corpus in Spanish + post-processing
5	0.756	0.795	0.775	Pre-processing + RoBERTa-base trained on Spanish general-domain corpus + Weak supervision + post-processing
24	0.779	0.769	0.774	Pre-processing and gazetteer lookup + Transformer-based ensemble
4	0.680	0.805	0.738	Data Augmentation + Joint Learning + post-processing
-	0.759	0.644	0.697	Terminology-based system
32	0.640	0.655	0.647	static and contextual embeddings with Flair NER framework.
16	0.836	0.494	0.621	Data Augmentation + BioBERT
-	0.505	0.625	0.559	-
28	0.504	0.461	0.481	Multilingual BERT
19	0.004*	0.004*	0.004*	Multilingual BERT
baseline	0.776	0.701	0.737	Terminology-based system

*Team 19 had problems in the evaluation phase due to the format used in the submission. The best system of team 51 was a post-workshop evaluation.

Table 13: Evaluation results for Task 10, Socialdisner: ranking with the best submission per team. Best result bolded, second best underlined. Metrics show micro-averaged F_1-score (F_1), precision (P), and recall (R)s.

performing teams opted for an architecture based on transformer models, often trained with heuristics chosen according to the characteristics of the task at hand, e.g. imbalance or texts not written in English. We note that among the 47 teams that submitted at least one predictions and wrote a paper description, 0.43% (20 teams) participated in multiple tasks, often with the same systems that were fine tuned to perform the different tasks. We believe this marks an important effort for the community to develop high-performing classifiers/label sequencers that are generalizable and reusable. The system description papers that are cited in Appendix 4 were each peer-reviewed by two reviewers and provide further details about the 47 teams' systems.

Acknowledgements

The work for #SMM4H 2022 at Cedars-Sinai Medical Center was supported by the National Institutes of Health (NIH) National Library of Medicine (NLM) [grant number R01LM011176]. The work on the set of tweets for Task 2 was done at Kazan Federal University and supported by the Russian Science Foundation [grant number 18-11-00284]. The work on Task 10-SocialDisNER has been supported by: the Encargo of PlanTL (SEDIA) to BSC; BIOMATDB, a European Union Horizon Europe coordination and support action under grant agreement no. 101058779; and the AI4ProfHealth project (PID2020-119266RA-I00)

The authors would also like to thank all those who reviewed the system description papers.

References

Omar Adjali, Fréjus A. A. Laleye, and Umang Aggarwal. 2022. Ofu@smm4h'22: Mining advent drug events using pretrained language models. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 171–175.

Mohammed Ali Al-Garadi, Sangmi Kim, Yuting Guo, Elise Warren, Yuan-Chi Yang, Sahithi Lakamana, and Abeed Sarker. 2022. Natural language model for automatic identification of intimate partner violence reports from twitter. *Array*, 15:100217.

Nestor Alvaro, Yusuke Miyao, Nigel Collier, et al. 2017. Twimed: Twitter and pubmed comparable corpus of drugs, diseases, symptoms, and their relations. *JMIR public health and surveillance*, 3(2):e6396.

Andrei-Marius Avram, Vasile Pais, and Maria Mitrofan. 2022. Racai@smm4h'22: Tweets disease mention detection using a neural lateral inhibitory mechanism. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 1–3.

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.

Francesco Barbieri, Luis Espinosa-Anke, and Jose Camacho-Collados. 2022. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. *Proceedings of the LREC, Marseille, France*, pages 20–25.

Alec Louis Clemente Candidato, Akshat Gupta, Xiaomo Liu, and Sameena Shah. 2022. Airjpmc@smm4h'22: Classifying self-reported intimate partner violence in tweets with multiple bert-based models. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 135–137.

José Canete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020:1–10.

Casimiro Pio Carrino, Joan Llop, Marc Pàmies, Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Joaquín Silveira-Ocampo, Alfonso Valencia, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. Pretrained biomedical language models for clinical nlp in spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 193–199.

Kendrick Cetina and Nuria García-Santa. 2022. Fre at socialdisner: Joint learning of language models for named entity recognition. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 68–70.

Mariia Chizhikova, Jaime Collado-Montañéz, Pilar López-Úbeda, Manuel C. Díaz-Galiano, L. Alfonso Ureña-López, and M. Teresa Martín-Valdivia. Sinai at clef 2022: Leveraging biomedical transformers to detect and normalize disease mentions. pages 265–273.

Mariia Chizhikova, Pilar López-Úbeda, Manuel C. Díaz-Galiano, L. Alfonso Ureña-López, and M. Teresa Martín-Valdivia. 2022. Sinai@smm4h'22: Transformers for biomedical social media text mining in spanish. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 27–30.

Tsz Hang Chu, Youzhen Su, Hanxiao Kong, Jingyuan Shi, and Xiaohui Wang. 2021. Online Social Support for Intimate Partner Violence Victims in China: Quantitative and Automatic Content Analysis. *Violence Against Women*, 27(3-4):339–358.

Daniel Claeser and Samantha Kent. 2022. Fraunhofer fkie @ smm4h 2022: System description for shared tasks 2, 4 and 9. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 103–107.

Jacob Cohen. 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

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. *arXiv preprint arXiv:1911.02116*.

Jaclyn D Cravens, Jason B Whiting, and Rola O Aamar. 2015. Why I Stayed/Left: An Analysis of Voices of Intimate Partner Violence on Social Media. *Contemporary Family Therapy*, 37(4):372–385.

Millon Das, Archit Mangrulkar, Ishan Manchanda, Manav Kapadnis, and Sohan Patnaik. 2022. Enolp musk@smm4h'22 : Leveraging pre-trained language models for stance and premise classification. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 156–159.

Shelby H. Davies, Miriam D. Langer, Ari Klein, Graciela Gonzalez-Hernandez, and Nadia Dowshen. 2022. Adolescent perceptions of menstruation on Twitter: Opportunities for advocacy and education. *Journal of Adolescent Health*, 71(1):94–104.

Vera Davydova and Elena Tutubalina. 2022. Smm4h 2022 task 2: Dataset for stance and premise detection in tweets about health mandates related to covid-19. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 216–220.

Alexander Philip Dawid and Allan M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Katherine Donegan, Hans Ovelgonne, Gavril Flores, Per Fuglerud, and Ada Georgescu. 2019. Social media and m-health data, subgroup report. *European Medicines Agency*.

Elissa S Epel, Alexandra D Crosswell, Stefanie E Mayer, Aric A Prather, George M Slavich, Eli Puterman, and Wendy Berry Mendes. 2018. More than a feeling: A unified view of stress measurement for population science. *Frontiers in neuroendocrinology*, 49:146–169.

Sumam Francis and Marie-Francine Moens. 2022. Kul@smm4h'22: Template augmented adaptive pre-training for tweet classification. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 153–155.

Raphael Antonius Frick and Martin Steinebach. 2022. Fraunhofer sit@smm4h'22: Learning to predict stances and premises in tweets related to covid-19 health orders using generative models. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 111–113.

Jia Fu, Sirui Li, Hui Ming Yuan, Zhucong Li, Zhen Gan, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. 2022. Casia@smm4h'22: A uniform health information mining system for multilingual social media texts. In *Proceedings of the Seventh Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, pages 143–147.

Luis Gasco, Darryl Estrada-Zavala, Eulàlia Farré-Maduell, Salvador Lima-López, Antonio Miranda-Escalada, and Martin Krallinger. 2022. Overview of the SocialDisNER shared task on detection of diseases mentions from healthcare related and patient generated social media content: methods, evaluation and corpora. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task.*

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in covid-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1596–1611.

Su Golder, Stephanie Chiuve, Davy Weissenbacher, Ari Klein, Karen O'Connor, Martin Bland, Murray Malin, Mondira Bhattacharya, Linda J. Scarazzini, and Graciela Gonzalez-Hernandez. 2019. Pharmacoepidemiologic evaluation of birth defects from health-related postings in social media during pregnancy. *Drug Safety*, 42(3):389–400.

Imane Guellil, Jinge Wu, Honghan Wu, Tony Sun, and Beatrice Alex. 2022. Edinburgh_ucl_health@smm4h'22: From glove to flair for handling imbalanced healthcare corpora related to adverse drug events, change in medication and self-reporting vaccination. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 148–152.

Asier Gutiérrez Fandiño, Jordi Armengol Estapé, Marc Pàmies, Joan Llop Palao, Joaquin Silveira Ocampo, Casimiro Pio Carrino, Carme Armentano Oller, Carlos Rodriguez Penagos, Aitor Gonzalez Agirre, and Marta Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.

Pan He, Chen YuZe, and Yanru Zhang. 2022. Zhegu@smm4h-2022: The pre-training tweet claim matching makes your prediction better. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 38–41.

Adrian Garcia Hernandez, Leung Wai Liu, Akshat Gupta, Vineeth Ravi, Saheed O. Obitayo, Xiaomo Liu, and Sameena Shah. 2022. Airjpmc@smm4h'22: Identifying self-reported spanish covid-19 symptom tweets through multiple-model ensembling. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 160–162.

Chenghao Huang, Xiaolu Chen, Yuxi Chen, Yutong Wu, Weimin Yuan, Yan Wang, and Yanru Zhang. 2022. zydhjh4593@smm4h'22: A generic pre-trained bert-based framework for social media health text classification. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 11–15.

Javier Huertas-Tato, Alejandro Martin, and David Camacho. 2022. Bertuit: Understanding spanish language in twitter through a native transformer. *arXiv preprint arXiv:2204.03465*.

Antonio Jimeno-Yepes, Andrew MacKinlay, Bo Han, and Qiang Chen. 2015. Identifying diseases, drugs, and symptoms in twitter. In *MEDINFO 2015: eHealth-enabled Health*, pages 643–647. IOS Press.

Keshav Kapur, Rajitha Harikrishnan, and Sanjay Singh. 2022. Manlp@smm4h'22: Bert for classification of twitter posts. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 42–43.

Akbar Karimi and Lucie Flek. 2022. Caisa@smm4h'22: Robust cross-lingual detection of disease mentions on social media with adversarial methods. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 168–170.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81.

Prabsimran Kaur, Guneet Singh Kohli, and Jatin Bedi. 2022. Arguably@smm4h'22: Classification of health related tweets using ensemble, zero-shot and fine-tuned language model. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 138–142.

Roshan Vivek Khatri, Sougata Saha, Souvik Das, and Rohini K Srihari. 2022. Ub health miners@smm4h'22: Exploring pre-processing techniques to classify tweets using transformer based pipelines. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 114–117.

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

Veysel Kocaman, Cabir Celik, Damla Gurbaz, Gursev Pirge, Bunyamin Polat, Halil Saglamlar, Meryem Vildan Sarikaya, Gokhan Turer, and David Talby. 2022. John_snow_labs@smm4h'22: Social media mining for health (#smm4h) with spark nlp. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 44–47.

Antoine Lain, Wonjin Yoon, Hyunjae Kim, Jaewoo Kang, and Ian Simpson. 2022. Ku_ed at socialdisner: Extracting disease mentions in tweets written in spanish. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 78–80.

Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. Clin-x: pre-trained language models and a study on cross-task transfer for concept extraction in the clinical domain. *arXiv preprint arXiv:2112.08754*.

Tzi-Mi Lin, Chao-Yi Chen, Yu-Wen Tzeng, and Lung-Hao Lee. 2022. Ncuee-nlp@smm4h'22: Classification of self-reported chronic stress on twitter using ensemble pre-trained transformer models. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 62–64.

Oscar William Lithgow-Serrano, Joseph Cornelius, Fabio Rinaldi, and Ljiljana Dolamic. 2022. mattica@smm4h'22: Leveraging sentiment for stance premise joint learning. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 75–77.

Leung Wai Liu, Akshat Gupta, Saheed Obitayo, Xiaomo Liu, and Sameena Shah. 2022a. Airjpmc@smm4h'22: Bert + ensembling = too cool: Using multiple bert models together for various covid-19 tweet identification tasks. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 163–167.

Xi Liu, Han Zhou, and Chang Su. 2022b. Pingantech at smm4h task1: Multiple pre-trained model approaches for adverse drug reactions. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 4–6.

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. 2021. 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. 2020. Deepademiner: A deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug effect mentions on twitter. *medRxiv*.

Heather L McCauley, Amy E Bonomi, Megan K Maas, Katherine W Bogen, and Teagen L O'Malley. 2018. # MaybeHeDoesntHitYou: Social Media Underscore the Realities of Intimate Partner Violence. *Journal of Women's Health*, 27(7):885–891.

Antonio Miranda-Escalada, Eulàlia Farré-Maduell, Salvador Lima-López, Luis Gascó, Vicent Briva-Iglesias, Marvin Agüero-Torales, and Martin Krallinger. 2021. The profner shared task on automatic recognition of occupation mentions in social media: systems, evaluation, guidelines, embeddings and corpora. In *Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*, pages 13–20.

Antonio Miranda-Escalada, Aitor Gonzalez-Agirre, Jordi Armengol-Estapé, and Martin Krallinger. 2020. Overview of automatic clinical coding: Annotations, guidelines, and solutions for non-english clinical cases at codiesp track of clef ehealth 2020. *CLEF (Working Notes)*, 2020.

Rosa M. Montañés-Salas, Irene López-Bosque, Luis García-Garcés, and Rafael del Hoyo-Alonso. 2022. Itainnova at socialdisner: A transformers cocktail for disease identification in social media in spanish. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 71–74.

Edgar Morais, José Luis Oliveira, Alina Trifan, and Olga Fajarda. 2022. Bioinfo@uavr@smm4h'22: Classification and extraction of adverse event mentions in tweets using transformer models. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 65–67.

Miguel Ortega-Martín, Alfonso Ardoiz, Jorge Álvarez, Oscar García-Sierra, and Adrián Alonso. 2022. dezzai@smm4h'22: Tasks 5 10 - hybrid models everywhere. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 7–10.

Christopher Palmer, Sedigheh Khademi, and Muhammad Javed. 2022. Chaai@smm4h'22: Roberta, gpt-2 and sampling - an interesting concoction. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 81–84.

Beatrice Portelli, Simone Scaboro, Emmanuele Chersoni, Enrico Santus, and Giuseppe Serra. 2022. Ailab-udine@smm4h'22: Limits of transformers and bert ensembles. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 130–134.

Vadim A. Porvatov and Natalia Semenova. 2022. Transformer-based classification of premise in tweets related to covid-19. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 108–110.

Matias Rojas, Jose Barros, Kinan R. Martin, Mauricio Araneda-Hernandez, and Jocelyn Dunstan. 2022. Pln cmm at socialdisner: Improving detection of disease mentions in tweets by using document-level features. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 52–54.

Vatsal Savaliya, Aakash Bhatnagar, Nidhir Bhavsar, and Muskaan Singh. 2022. Innovators@smm4h'22: An ensembles approach for stance and premise classification of covid-19 health mandates tweets. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 126–129.

Sanja Scepanovic, Enrique Martin-Lopez, Daniele Quercia, and Khan Baykaner. 2020. Extracting medical entities from social media. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 170–181.

Ana Lucía Schmidt, Raul Rodriguez-Esteban, Juergen Gottowik, and Mathias Leddin. 2022. Applications of quantitative social media listening to patient-centric drug development. *Drug Discovery Today*.

Isabel Segura-Bedmar, Ricardo Revert, and Paloma Martínez. 2014. Detecting drugs and adverse events from spanish social media streams. In *Proceedings of the 5th international workshop on health text mining and information analysis (LOUHI)*, pages 106–115.

Aman Sinha, Cristina Garcia Holgado, Marianne Clausel, and Matthieu Constant. 2022. Iai @ socialdisner : Catch me if you can! capturing complex disease mentions in tweets. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 85–89.

Sharon Smith, Xinjian Zhang, Kathleen Basile, Melissa Merrick, Jing Wang, Marcie-jo Kresnow, and Jieru Chen. 2018. The National Intimate Partner and Sexual Violence Survey: 2015 Data Brief — Updated Release. Technical report, National Center for Injury Prevention and Control, Centers for Disease Control and Prevention, Atlanta, Georgia.

Afrin Sultana, Nihad Karim Chowdhury, and Abu Nowshed Chy. 2022. Csecu-dsg@smm4h'22: Transformer based unified approach for classification of changes in medication treatments in tweets and webmd reviews. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 118–122.

Antonio Tamayo, Diego Burgos, and Alexander Gelbukh. 2022. Nlp-cic-wfu at socialdisner: Disease mention extraction in spanish tweets using transfer learning and search by propagation. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 19–22.

Ramya Tekumalla and Juan M Banda. 2021. Using weak supervision to generate training datasets from social media data: a proof of concept to identify drug mentions. *Neural Comput. Appl.*, pages 1–9.

Atnafu Lambebo Tonja, Olumide Ebenezer Ojo, Mohammed Arif, Abdul Gafar Manuel Meque, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2022. Cic nlp at smm4h 2022: a bert-based approach for classification of social media forum posts. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 58–61.

Paul Trust, Rosane Minghim, Ahmed Zahran, Provia Kadusabe, and Kizito Omala. 2022. Ucc-nlp@smm4h'22:label distribution aware long-tailed learning with post-hoc posterior calibration applied to text classification. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 90–94.

Gökçe Uludoğan and Zeynep Yirmibeşoğlu. 2022. Boun-tabi@smm4h'22: Text-to-text adverse drug event extraction with data balancing and prompting. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 31–34.

Reshma Unnikrishnan, Kamath S. Sowmya, and V. S. Ananthanarayana. 2022. Halelab_nitk@smm4h'22: Adaptive learning model for effective detection, extraction and normalization of adverse drug events from social media data. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 95–97.

U.S. Food and Drug Administration. 2020. *Patient-Focused Drug Development: Collecting Comprehensive and Representative Input Guidance for Industry, Food and Drug Administration Staff,and Other Stakeholders.*

Gary R VandenBos. 2007. *APA dictionary of psychology.* American Psychological Association.

Harsh Verma, Parsa Bagherzadeh, and Sabine Bergler. 2022. Claclab at socialdisner: Using medical gazetteers for named-entity recognition of disease mentions in spanish tweets. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 55–57.

Chunchen Wei, Ran Bi, and Yanru Zhang. 2022. uestcc@smm4h'22: Roberta based adverse drug events classification on tweets. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 35–37.

Davy Weissenbacher, Suyu Ge, Ari Klein, Karen O'Connor, Robert Gross, Sean Hennessy, and Graciela Gonzalez-Hernandez. 2021. Active neural networks to detect mentions of changes to medication treatment in social media. *Journal of the American Medical Informatics Association*, 28(12):2551–2561.

Lynn Westbrook. 2015. Intimate Partner Violence Online: Expectations and Agency in Question and Answer Websites. *Journal of the Association for Information Science and Technology*, 66.

Orest Xherija and Hojoon Choi. 2022. Complx@smm4h'22: In-domain pretrained language models for detection of adverse drug reaction mentions in english tweets. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 176–181.

Huabin Yang, Zhongjian Zhang, and Yanru Zhang. 2022a. yiriyou@smm4h'22: Stance and premise classification in domain specific tweets with dual-view attention neural networks. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 23–26.

Yuan-Chi Yang, Angel Xie, Sangmi Kim, Jessica Hair, Mohammed Ali Al-Garadi, and Abeed Sarker. 2022b. Automatic detection of Twitter users who express chronic stress experiences via supervised machine learning and natural language processing. *Computers, Informatics, Nursing*, Article in press.

Antonio Jimeno Yepes and Karin Verspoor. 2022. Readbiomed@socialdisner: Adaptation of an annotation system to spanish tweets. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 48–51.

Sourabh Satish Zanwar, Daniel Wiechmann, Yu Qiao, and Elma Kerz. 2022. Mantis at smm4h'2022: Pretrained language models meet a suite of psycholinguistic features for the detection of self-reported chronic stress. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 16–18.

Yan Zhuang and Yanru Zhang. 2022. Yet@smm4h'22: Improved bert-based classification models with rdrop and polyloss. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 98–102.

Mohammad Zohair, Nidhir Bhavsar, Aakash Bhatnagar, and Muskaan Singh. 2022. Innovators @ smm4h'22: An ensembles approach for self-reporting of covid-19 vaccination status tweets. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 123–125.

Appendix A. Team numbers and System Description Papers

#Team	System description paper
1	(Adjali et al., 2022)
2	(Avram et al., 2022)
3	(Candidato et al., 2022)
4	(Cetina and García-Santa, 2022)
5	(Chizhikova et al., 2022)
6	(Claeser and Kent, 2022)
7	(Das et al., 2022)
8	(Francis and Moens, 2022)
9	(Frick and Steinebach, 2022)
10	(Fu et al., 2022)
11	(Guellil et al., 2022)
12	(He et al., 2022)
13	(Hernandez et al., 2022)
14	(Huang et al., 2022)
15	(Kapur et al., 2022)
16	(Karimi and Flek, 2022)
17	(Kaur et al., 2022)
18	(Khatri et al., 2022)
19	(Kocaman et al., 2022)
20	(Lain et al., 2022)
21	(Lin et al., 2022)
22	(Lithgow-Serrano et al., 2022)
23	(Liu et al., 2022a)
24	(Montañés-Salas et al., 2022)
25	(Morais et al., 2022)
26	(Ortega-Martín et al., 2022)
27	(Palmer et al., 2022)
28	(Portelli et al., 2022)
29	(Porvatov and Semenova, 2022)
30	(Rojas et al., 2022)
31	(Savaliya et al., 2022)
32	(Sinha et al., 2022)
33	(Sultana et al., 2022)
34	(Tamayo et al., 2022)
35	(Tonja et al., 2022)
36	(Trust et al., 2022)
37	(Uludoğan and Yirmibeşoğlu, 2022)
38	(Unnikrishnan et al., 2022)
39	(Verma et al., 2022)
40	(Wei et al., 2022)
41	(Xherija and Choi, 2022)
42	(Yang et al., 2022a)
43	(Yepes and Verspoor, 2022)
44	(Zanwar et al., 2022)
45	(Liu et al., 2022b)
46	(Zhuang and Zhang, 2022)
47	(Zohair et al., 2022)