Overview of the Sixth Social Media Mining for Health Applications (#SMM4H) Shared Tasks at NAACL 2021

Arjun Magge
University of Pennsylvania
Philadelphia, PA, USA

Ari Z. Klein
University of Pennsylvania
Philadelphia, PA, USA

Antonio Miranda-Escalada
Barcelona Supercomputing Center
Barcelona, Spain

Mohammed Ali Al-garadi
Emory University
Atlanta, GA, USA

Ilseyar Alimova
Kazan Federal University
Kazan, Russia

Zulfat Miftahutdinov
Kazan Federal University
Kazan, Russia

Eulàlia Farré-Maduell
Barcelona Supercomputing Center
Barcelona, Spain

Salvador Lima López
Barcelona Supercomputing Center
Barcelona, Spain

Ivan Flores
University of Pennsylvania
Philadelphia, PA, USA

Karen O'Connor
University of Pennsylvania
Philadelphia, PA, USA

Davy Weissenbacher
University of Pennsylvania
Philadelphia, PA, USA

Elena Tutubalina
Kazan Federal University
Kazan, Russia

Abeed Sarker
Emory University
Atlanta, GA, USA

Juan M. Banda
Georgia State University
Atlanta, Georgia

Martin Krallinger
Barcelona Supercomputing Center
Barcelona, Spain

Graciela Gonzalez-Hernandez
University of Pennsylvania
Philadelphia, PA, USA

{arjun.magge, ariklein, ivan.flores, karoc, dweissen, gragon}@pennmedicine.upenn.edu
{antonio.miranda, eulalia.farre, salvador.limalopez, martin.krallinger}@bsc.es
{alimovailseyar, zulfatmi, tutubalinaev}@gmail.com
{m.a.al-garadi, abeed}@dbmi.emory.edu
jbanda@gsu.edu

Abstract

The global growth of social media usage over the past decade has opened research avenues for mining health related information that can ultimately be used to improve public health. The Social Media Mining for Health Research and Applications (#SMM4H) shared tasks in its sixth iteration sought to advance the use of social media texts such as Twitter for pharmacovigilance, disease tracking and patient centered outcomes. #SMM4H 2021 hosted a total of eight tasks that included reruns of adverse drug effect extraction in English and Russian and newer tasks such as detecting medication non-adherence from Twitter and WebMD forum, detecting self-reported adverse pregnancy outcomes, detecting cases and symptoms of COVID-19, identifying occupations mentioned in Spanish by Twitter users, and detecting self-reported breast cancer diagnosis. The eight tasks included a total of 12 individual subtasks spanning three languages requiring methods for binary classification, multiclass classification, named entity recognition (NER) and entity normalization. With a total of 97 registering teams and 40 teams submitting predictions, the interest in the shared tasks grew by 70% and participation grew by 38% compared to the previous iteration.

1 Introduction

The Social Media Mining for Health (#SMM4H) shared tasks aim to foster community participation in tackling natural language processing (NLP) challenges in social media texts for health applications. The tasks hosted annually attract newer methods for extraction of meaningful health related infor-

Proceedings of the Sixth Social Media Mining for Health Workshop 2021, pages 21–32

mation from noisy social media sources such as Twitter and WebMD where the information of interest is often sparse and noisy. The NLP methods required for the eight tasks spanned the categories of text classification, named entity recognition and entity normalization. Systems developed for the tasks often require the use of NLP techniques such as noise removal, class weighting, undersampling, oversampling, multi-task learning, transfer learning and semi-supervised learning to improve over traditional methods.

The sixth iteration of #SMM4H hosted eight tasks with a total of twelve individual subtasks. Similar to previous years, the most tasks centered around pharmacovigilance (i.e. ADE extraction, medication adherence) and patient centered outcomes (i.e. adverse pregnancy outcomes, breast cancer diagnosis). This year the shared tasks featured the addition of COVID-19 related tasks such as detection of self reported cases of COVID-19 and symptoms of COVID-19, as well as extraction of professions and occupations for the purposes of risk analysis. The individual tasks are listed below:

1. Classification, extraction and normalization of adverse drug effect (ADE) mentions in English tweets
   (a) Classification of tweets containing ADEs
   (b) Span extraction of ADE mentions
   (c) Span extraction and normalization of ADE mentions
2. Classification of Russian tweets for detecting presence of ADE mentions
3. Classification of change in medications regimen on
   (a) Twitter
   (b) WebMD
4. Classification of tweets self-reporting adverse pregnancy outcomes
5. Classification of tweets self-reporting potential cases of COVID-19
6. Classification of COVID-19 tweets containing symptoms
7. Identification of professions and occupations (ProfNER) in Spanish tweets
   (a) Classification of tweets containing mentions of professions and occupations
   (b) Span extraction of professions and occupations
8. Classification of self-reported breast cancer posts on Twitter

Teams interested in participating were allowed to register for one or more tasks/subtasks. On successful registration, teams were provided with annotated training and validation sets of tweets for each task. In total, 97 teams registered for one or more tasks. The annotated datasets contained examples of input text and output labels which the participants could use to train their methods. During the final evaluation period which lasted four days for each task, teams were provided with a evaluation datasets which contained only the input texts. Participants were required to submit label predictions for the input texts which would be evaluated against the annotated labels. The submissions were facilitated through Codalab [1] and participants were allowed to make up to two prediction submissions for each of the subtasks. Of the 97 registered teams, 40 teams submitted one or more predictions towards the shared tasks.

The remainder of the document is as follows, in Section 2, we briefly describe the individual task objectives and research challenges associated with them. In Section 3, we present the evaluation results and a brief summary of each team's best-performing system for each subtask. Appendix A provides the system description papers corresponding to the team numbers.

2 Tasks

2.1 Task 1: Classification, extraction and normalization of ADE mentions in English tweets

The objectives of Task 1 was to develop automated methods to extract adverse drug effects from tweets containing drug mentions for social media pharmacovigilance. Task 1 and their subtasks have been the longest running tasks at SMM4H. This task presented three challenges listed as subtasks in increasing order of complexity wherein in the systems developed must contain one or more components to accomplish the following: (Task 1a) Classify tweets that contain one or more adverse effects (AE) or also known as adverse drug effect (ADE), (Task 1b) Classify the tweets containing ADEs from Task 1a and further extract the text span of reported ADEs in tweets, and (Task 1c) Classify the tweets containing ADE, extract the text span and further normalize these colloquial mentions to their standard concept IDs in the MedDRA ontology's preferred terms.

[1] https://competitions.codalab.org/

The training dataset contains a total of 18,300 tweets with 17,385 tweets for training, 915 tweets for validation (Magge et al., 2020). Participants were allowed to use both training and validation set for training their models for the evaluation stage. The evaluation was performed on 10,984 tweets. The tweets were manually annotated at three levels corresponding to the three subtasks: (a) tweets that contained one or more mentions of ADE had the *ADE* label assigned to them, (b) each ADE was annotated with the starting and ending indices of the ADE mention in the text, and (c) each ADE also contained the normalized MedDRA lower-level term (LLT) that were evaluated at the higher preferred term (PT) level. There are more than 79,000 MedDRA LLT terms and more than 23,000 preferred terms in the MedDRA ontology. The combined test and training dataset contains 2,765 ADE annotations with 669 unique LLT identifiers. The test set contained 257 LLT terms that were not part of the training set, making it important for the developed system to be capable of extracting ADEs that were not part of the training set. While subtasks 1a and 1b presented a class imbalance problem wherein the classification task needs to take into account that only around 7% of the tweets contain ADEs, subtask 1c presented a challenge with the large potential label space. Systems were evaluated and ranked based on the F_1-score for the *ADE* class, overlapping ADE mentions and overlapping ADE mentions with matching PT ids for subtasks 1a, 1b and 1c respectively.

2.2 Task 2: Classification of Russian tweets for detecting presence of ADE mentions

Task 2 presented a similar challenge to Task 1a wherein the designed system is capable of identifying tweets in Russian that contain one or more adverse drug effects. The dataset contains 11,610 tweets for training and validation, with 1073 (9.24%) tweets that report an ADE. The test set contains 9095 tweets, with 778 (8.55%) tweets that report an ADE. All of the Russian tweets were dual annotated; first, three *Yandex.Toloka*[2] annotators' crowd-sourced labels were aggregated into a single label (Dawid and Skene, 1979), and then the tweets were labeled by a second annotator from KFU. Inter-annotator agreement was 0.74 (Cohen's kappa). Systems were evaluated based on the F_1-score for the "positive" class (i.e., tweets that report an adverse effect).

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

2.3 Task 3: Classification of change in medications regimen in tweets

Task 3 is a binary classification task that involves distinguishing social media posts where users self-declare changing their medication treatments, regardless of being advised by a health care professional to do so. Posts with self-declaration of changes are annotated as "1", other posts are annotated as "0". Such changes are, for example, not filling a prescription, stopping a treatment, changing a dosage, forgetting to take the drugs, etc. This task is the first step toward detecting patients non-adherent to their treatments and their reasons on social media. The data consists of two corpora: 9,830 tweets from Twitter and 12,972 drug reviews from WebMD. Positive and negative tweets are naturally imbalanced with a 10.38 Imbalance Ratio whereas negative and positive WebMD reviews are naturally balanced with a 0.80 Imbalance Ratio. Each corpus is split into a training (5,898 Tweets / 10,378 Reviews), a validation (1,572 Tweets / 1,297 Reviews), and a test subset (2,360 Tweets / 1,297 Reviews). We provided to the participants the training and validation subsets for both corpora and we evaluated on both test subsets independently. We added in the test sets additional reviews and tweets as decoys to avoid manual corrections of the predicted labels. We evaluated participants' systems based on the F_1-score for the "positive" class (i.e., tweets or reviews mentioning a change in medication treatments).

2.4 Task 4: Classification of tweets self-reporting adverse pregnancy outcomes

Despite the prevalence of miscarriage, stillbirth, preterm birth, and low birthweight, their causes remain largely unknown. To enable the use of Twitter data as a complementary resource for epidemiology of these adverse pregnancy outcomes, Task 4 is a binary classification task that involves automatically distinguishing tweets that potentially report a personal experience of an adverse pregnancy outcome ("outcome" tweets) from those that do not ("non-outcome" tweets). The training set (Klein and Gonzalez-Hernandez, 2020) contains 6487 annotated tweets: 3653 (45%) "outcome" tweets (annotated as "1") and 4456 (55%) "non-outcome" tweets (annotated as "0"). The test set contains 1622 annotated tweets: 731 (45%) "outcome" tweets and

891 (55%) "non-outcome" tweets. Inter-annotator agreement (Cohen's kappa) was 0.90. Systems were evaluated based on the F_1-score for the "outcome" class.

2.5 Task 5: Classification of tweets self-reporting potential cases of COVID-19

The COVID-19 pandemic has presented challenges for actively monitoring its spread based on testing alone. Task 5 is a binary classification task that involves automatically distinguishing tweets that self-report potential cases of COVID-19 ("potential case" tweets) from those that do not ("other" tweets), where "potential case" tweets broadly include those indicating that the user or a member of the user's household was denied testing for COVID-19, showing symptoms of COVID-19, potentially exposed to cases of COVID-19, or had had experiences that pose a higher risk of exposure to COVID-19. The training set (Klein et al., 2021) contains 7181 tweets: 1148 (16%) "potential case" tweets (annotated as "1") and 6033 (84%) "other" tweets (annotated as "0"). The test set contains 1795 annotated tweets: 308 (17%) "potential case" tweets and 1487 (83%) "other" tweets. Inter-annotator agreement (Cohen's kappa) was 0.77. Systems were evaluated based on the F_1-score for the "potential case" class.

2.6 Task 6: Classification of COVID-19 tweets containing symptoms

Identifying personal mentions of COVID-19 symptoms requires distinguishing personal mentions from other mentions such as symptoms reported by others and references to news articles or other sources. The classification medical symptoms from COVID-19 Twitter posts presents two key issues: First, there is plenty of discourse around news and scientific articles that describe medical symptoms. While this discourse is not related to any user in particular, it enhances the difficulty of identifying valuable user-reported information. Second, many users describe symptoms that other people experience, instead of their own, as they are usually caregivers or relatives of people presenting the symptoms. This makes the task of separating what the user is self-reporting particularly tricky, as the discourse is not only around personal experiences. Task 6 is considered a three-way classification task where the target classes are: (1) self-reports, (2) non-personal reports, and (3) literature/news mentions. In this task, the tweets were sampled from the collections created by Banda et al. (2020b). The sampled tweets were manually annotated by clinicians for extracting long-term patient-reported symptoms of COVID-19 Banda et al. (2020a). The annotated dataset contained a total of 16,067 tweets, 9567 of which were used for training and 6500 used for testing. Systems were evaluated and ranked based on micro-F_1-scores.

2.7 Task 7: Identification of professions and occupations in Spanish tweets (ProfNER)

Extraction of occupations from health-related content is critical for planning public health measures and epidemiological surveillance systems not only in the context of infectious disease outbreaks like COVID-19. Here, occupations refer to paid (profession) and unpaid (activity) working activities, as well as working status such as "student" or "retired". Occupational risks due to exposure to infectious/hazardous agents or mental health conditions linked to occupational stress require systematic extraction of professions from different types of content including user generate contents like social media. Task 7 focused on the detection of occupations from COVID-related tweets in Spanish (the ProfNER corpus[3]). The aim was to enable detection of health-related issues linked to occupations, with special emphasis on the COVID-19 pandemic. In subtask 7a (text classification), participants had to classify tweets containing occupation mentions in Spanish COVID-related tweets and in subtask 7b (named entity recognition), required extraction of text spans mentioning occupations.

This task presents multiple challenges. The classification task had to cope with class imbalance issues, as only 23.3% of the provided tweets mentioned occupations. Secondly, the occupation mention detection required advanced named entity recognition approaches to deal with the heterogeneity and colloquial ways people were referring to occupations in social media. In both subtasks, participating systems had to process noisy user-generated text in Spanish and scale up to a large number of records. For subtask 7a, systems were evaluated and ranked based on the F_1-score for the positive class i.e. tweets containing an occupation mention and for subtask 7b, F_1-score for the *PROFESSION* and *SITUACION_LABORAL* classes where the spans overlap entirely was used.

[3] https://doi.org/10.5281/zenodo.4309356

2.8 Task 8: Classification of self-reported breast cancer posts on Twitter

Breast cancer patients often discontinue their long-term treatments, such as hormone therapy, increasing the risk of cancer recurrence. These discontinuations may be caused by adverse patient-centered outcomes (PCOs) due to hormonal drug side effects or other factors. PCOs are not detectable through laboratory tests and are sparsely documented in electronic health records. Thus, there is a need to explore complementary sources of information for PCOs associated with breast cancer treatments. Social media is a promising resource but extracting true PCOs from it requires the accurate detection of self-reported breast cancer patients. Task 8 focused on developing systems for this first step i.e. identifying tweets with self-reported breast cancer diagnosis. The dataset for Task 8 contained a total of 3815 tweets for training and 1204 tweets for testing. In this task, only about 26% of the tweets contains such self-reports (S) and 74% of the tweets are non-relevant (NR). Systems designed for this task need to automatically identify tweets in the self-reports category. Systems were evaluated based on the F_1-score for the self-reports (S) class.

3 Results

3.1 Task 1: Classification, extraction and normalization of ADE mentions in English tweets

Table 1 presents the results from Task 1. The best performance achieved in task 1a was an F_1-score of 0.61 which initially appears to be 3 percentage points (p.p) lower than previous year's score of 0.64. However on closer examination we find that in addition to the datasets being different, participants in SMM4H 2020 used additional corpora to train their systems. The best performance in ADE extraction i.e. task 1b was an F_1-score of 0.51 which used multi-task learning methods to optimize their models across classification and the NER task. For both tasks 1a and 1b, we find that the systems with the best *Recall* scores ranked the best among all submissions emphasizing the importance of developing systems that account for the class imbalance. The best performance for the overall task of ADE extraction and normalization i.e. task 1c was 0.29 which was achieved by leveraging annotations from other datasets and incorporating semi-supervised learning across corpora similar to previous year's leading system (Miftahutdinov et al., 2020). Overall, the percentage of teams using transformer architectures for subtasks 1a and 1c rose from 80% in SMM4H 2020 to 100% in SMM4H 2021.

3.2 Task 2: Classification of Russian tweets for detecting presence of ADE mentions

In total, 30 teams were registered and 3 teams submitted models' predictions during the evaluation period. Table 2 presents the F_1-score, precision and recall for the ADE class, for each of the teams' best-performing systems and two baselines for Task 2. Compared to last year's results for this task, arithmetic median of all submissions made by teams increased from 0.42 F_1 to 0.51 F_1. Two best-performing systems for this task in #SMM4H 2020 (Klein et al., 2020a) achieved an F_1-score of 0.51 (Gusev et al., 2020; Miftahutdinov et al., 2020), while the best-performing system in #SMM4H 2021 achieved an F_1-score of 0.57. All teams used a transformer-based architecture.

3.3 Task 3: Classification of change in medications regimen in tweets

Despite the interest for task 3, with 29 teams registered, only one team submitted their predictions during the evaluation period. We reported the performances achieved by the best baseline classifiers and the best team's classifiers in Table 3. The leading team chose a standard architecture for their classifier: a transformer encoder followed by an average pooling layer, a linear layer, and a softmax layer for the prediction. They focused on the impact of the corpora used to pre-train two transformers models, BERT and RoBERTa. They evaluated single and ensemble models pre-trained on corpora of different genres and domains - tweets, clinical notes/biomedical research articles, or Wikipedia. While the ensemble of transformers did not improve on the performance of the default BERT-base model used by the baseline on the WebMD corpus, it proves to be beneficial on the imbalanced Twitter corpus. The baseline classifier handles the imbalance of the Twitter corpus by pre-training with active learning a CNN on the WebMD corpus to transfer the knowledge learned on this balanced corpus(Weissenbacher et al., 2020). The team used more successfully a conventional approach by oversampling the positive tweets of the training set and the ensemble of the predictions of several transformer models. These strategies are not exclusive

Task	Team	F_1	P	R	System Summary
Task 1a	4	**0.61**	0.515	**0.752**	RoBERTa undersampling and oversampling
	23	**0.61**	0.552	0.681	RoBERTa + ChemBERTa
	12	0.54	**0.603**	0.489	BERT variants with oversampling and ensemble
	16	0.49	0.592	0.417	BERTweet with automatically curated (pseudo) data
	10	0.46	0.472	0.456	BERT trained with class weights
	20	0.46	0.523	0.409	BERTweet with class weights
	26	0.44	0.491	0.393	Multi-task learning model with BioBERT and class weights
	27	0.40	0.405	0.401	RoBERTa with SMOTE and data augmentation
	22	0.40	0.521	0.327	BERT ensembles with oversampling
	-	0.31	0.500	0.222	-
	24	0.23	0.135	0.726	BERT
Task 1b	26	**0.51**	0.514	**0.514**	Multi-task learning with selective oversampling
	10	0.50	0.555	0.459	RoBERTa with FastText and byte-pair embeddings
	4	0.50	0.493	0.505	RoBERTa
	22	0.49	**0.681**	0.385	-
	16	0.42	0.381	0.475	BERT with BiLSTM+CRF layer
	23	0.40	0.420	0.382	EnDR-BERT with data from CADEC and COMETA corpora
	24	0.37	0.580	0.275	BERT with joint NER and Normalization
Task 1c	23	**0.29**	0.301	0.275	EnDR-BERT with data from CADEC and COMETA corpora
	28	0.24	0.317	0.196	ELMO, CharCNN and Glove in trained jointly
	24	0.24	**0.371**	0.178	BERT with joint NER and Normalization
	16	0.2	0.139	**0.342**	BERTweet and similarity measures
	26	0.16	0.160	0.170	Multi-task learning with selective oversampling

Table 1: Evaluation results for Task 1:Classification, extraction and normalization 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
23	**0.57**	**0.58**	0.57	EnRuDR-BERT + ChemBERTa
-	0.54	0.57	0.52	-
3	0.47	0.39	**0.59**	BERT-based model trained additionally on augmented texts
Baseline #1	0.41	0.40	0.42	CNN-based classifier with FastText embeddings
Baseline #2	0.50	0.45	0.56	BERT-based classifier

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

Task	Team	F_1	P	R	System Summary
Task 3a	22	**0.68**	**0.72**	**0.64**	Ensemble of 3 transformers-based models trained with oversampling
	Baseline	0.50	0.47	0.53	CNN trained with transfer and active learning
Task 3b	22	0.86	0.84	**0.89**	Ensemble of 4 transformer-based models
	Baseline	**0.87**	**0.87**	0.88	BERT-based

Table 3: Evaluation results for Task 3: Detecting change in medication treatment in tweets (Task 3a) and WebMD reviews (Task 3b). Metrics show F_1-scores (F_1), precision (P), and recall (R) for the *positive* class.

Team	F_1	P	R	System Summary
22	**0.93**	**0.94**	0.92	BERTweet, RoBERTa-Large
15	0.93	0.91	**0.95**	RoBERTa-Large, language model and classifier fine-tuning
27	0.92	0.89	**0.95**	RoBERTa
-	0.78	0.78	0.78	-

Table 4: Evaluation results for Task 4: Classification of tweets self-reporting adverse pregnancy outcomes. Metrics show F_1-scores (F_1), precision (P), and recall (R) for the *positive* class.

to each other and could be used in a common classifier for future work.

3.4 Task 4: Classification of tweets self-reporting adverse pregnancy outcomes

Table 4 presents the precision, recall, and F_1-score for the *outcome* class, for each of the four team's best-performing system for Task 4. The three top-performing systems achieved similar F_1-scores using RoBERTa pre-trained transformer models. The leading team achieved the marginally highest F_1-score (0.93) using an ensemble of RoBERTa and BERTweet pre-trained models. While the leading team also achieved the highest precision (0.94), the highest recall (0.95) was achieved by another team using the RoBERTa model alone. Overall, using a model pre-trained on tweets did not significantly improve performance for this task. The RoBERTa-based classifiers outperformed a BERT-based classifier (F_1-score = 0.88) presented in recent work (Klein et al., 2020b).

3.5 Task 5: Classification of tweets self-reporting potential cases of COVID-19

Table 5 presents the precision, recall, and F_1-score for the "potential case" class, for each of the 14 team's best-performing system for Task 5. The team with the highest performance F_1-score (0.79), precision (0.78), and recall (0.79) used an ensemble of five BERT-based pre-trained transformer models, including models pre-trained on tweets related to COVID-19. To address the class imbalance, the leading team over-sampled the "potential case" class, and further augmented the "potential case" class using paraphrasing via round-trip translation from English into German, and then back into English. Teams placing second and third achieved F_1-scores of 0.77 and 0.76, respectively, using COVID-Twitter-BERT, while the teams (that submitted system descriptions) that achieved F_1-scores of less than 0.76 did not use models pre-trained on tweets related to COVID-19. The leading team outperformed a benchmark classifier presented in recent work (Klein et al., 2021), which was based on COVID-Twitter-BERT and achieved an F_1-score (0.76) similar to that of the teams placing second and third.

3.6 Task 6: Classification of COVID-19 tweets containing symptoms

Table 6 presents the precision, recall, and F_1-score for Task 6. Unsurprisingly 7 out of the top 11 submissions used BERT or variations of it. Some teams fine tuned their models with additional COVID-19 Twitter data. The best performing team used a fine-tuned version of CT-BERT, achieving a 0.95 F1-score. While most models used are more complex deep learning architectures, team 7 managed to score higher than the median submission scores with a less complex multi-layer perceptron classifier. We believe the high scores in this task were due to the somewhat well-balanced dataset provided without the large class imbalance usually seen in Twitter data.

3.7 Task 7: Identification of professions and occupations in Spanish tweets (ProfNER)

Table 7 presents the Tweet Classification (subtask 7b) and Named Entity Recognition results (subtask 7b). In subtasks 7a and 7b, best-performing systems have effectively combined contextual embeddings or language models with the popular architecture RNN-CRF. For instance, the Recognai team has won both subtasks integrating the pre-trained Spanish language model BETO (Cañete et al., 2020) with an RNN-CRF engine built on top of the FastText medical embeddings (Soares et al., 2019). Besides, lighter models have usually been complemented with gazetteers either built from the training data or gathered from popular occupational terminologies.

3.8 Task 8: Classification of self-reported breast cancer posts on Twitter

Table 8 presents the F_1-score, precision and recall for the *self-reports* class (detection of self-reported breast cancer patient) for the participating teams. The leading team achieved a performance of F_1-score of 0.87. The leading team pre-processed the texts by tokenizing and normalizing tokens by replacing URLs with special tokens and replacing emojis with their semantic expressions. The leading team used BERTweet to encode tweet text and make a binary prediction according to the corresponding pooling vector. The analysis of the results shows that almost all top perform teams have achieved similar/comparable precision. However, the best performing team's recall was 5 p.p higher than the other teams which led to overall improve-

Team	F_1	P	R	System Summary
12	**0.79**	**0.78**	**0.79**	BERT ensemble, oversampling, data augmentation
8	0.77	0.77	0.78	COVID-Twitter-BERT
-	0.77	0.77	0.77	-
-	0.76	0.77	0.76	-
-	0.76	0.73	0.78	-
13	0.76	0.78	0.73	COVID-Twitter-BERT, Twitter-RoBERTa-Base, RoBERTa-Large
15	0.75	0.75	0.76	RoBERTa-Large, Task 6 corpus
22	0.75	0.73	0.77	RoBERTa-Base, RoBERTa-Large, BERTweet
25	0.72	0.70	0.73	XLNet, data augmentation
-	0.70	0.74	0.67	-
6	0.67	0.68	0.65	DistillBERT
29	0.53	0.74	0.41	BERT
-	0.43	0.47	0.39	-
-	0.36	0.56	0.26	-

Table 5: Evaluation results for Task 5: Classification of tweets self-reporting potential cases of COVID-19. Metrics show F_1-scores (F_1), precision (P), and recall (R) for the *positive* class.

Team	F_1	P	R	System Summary
8	**0.95**	**0.9477**	**0.9477**	**Fine-Tuned CT-BERT**
22	0.94	0.9449	0.9449	BERT + RoBERTa Large
11	0.94	0.9448	0.9448	XLNet
12	0.94	0.944	0.944	BERT-base ensemble model with data cleaning & domain specific BERT
4	0.94	0.9411	0.9411	BERTweet + 23 million COVID-19 tweets
24	0.94	0.9406	0.9406	Fine-Tuned BERT-base model
7	0.93	0.9337	0.9337	ML Model - multi-layer perceptron classifier
9	0.93	0.9325	0.9325	Fine-Tuned BERT with small-BERT pre-processing
29	0.84	0.8415	0.8415	BiLSTM
6	0.4	0.3951	0.3951	DistilBERT

Table 6: Evaluation results for Task 6: Classification of COVID-19 tweets containing symptoms. Metrics show micro-F_1-scores (F_1), precision (P), and recall (R)

Task	Team	F_1	P	R	System Summary
Task 7a	17	**0.93**	0.93	**0.93**	BETO language model and RNN with pre-trained word vectors
	2	0.92	0.92	0.92	Language model fine-tuned with ProfNER training corpus
	21	0.92	**0.95**	0.89	Data augmentation, BiLSTM-CRF with FLAIR and FastText embeddings
	10	0.90	**0.95**	0.86	Contextualized embeddings with BiLSTM-CRF
	1	0.89	0.89	0.88	NeuroNER with diverse word embeddings and a gazetteer
	-	0.85	0.93	0.78	-
	19	0.83	0.92	0.76	mBERT fine-tuned on augmented data using back-translation
	14	0.59	0.76	0.48	GloVe embeddings, BiLSTM and an occupations dictionary
	Baseline	0.8	0.75	0.87	Levenshtein distance to find mentions from training set
Task 7b	17	**0.84**	0.84	**0.84**	BETO language model and RNN with pre-trained word vectors
	21	0.73	0.81	0.66	Data augmentation, BiLSTM-CRF with FLAIR and FastText embeddings
	10	0.82	0.85	0.79	Contextualized embeddings with BiLSTM-CRF
	1	0.76	0.78	0.74	NeuroNER with diverse word embeddings
	14	0.73	0.82	0.65	CRF with an occupations dictionary
	5	0.82	**0.88**	0.77	Encoder-decoder architecture with attention feeded by different embeddings
	16	0.73	0.86	0.64	Voting of 30 models with Spanish BERT and BiLSTM modules
	Baseline	0.40	0.36	0.44	Levenshtein distance to find mentions from training set

Table 7: Evaluation results for Task 7: Identification of professions and occupations in Spanish tweets (ProfNER). Metrics show F_1-scores (F_1), precision (P), and recall (R) for the *positive* class on task 7a and micro averaged *PROFESSION* and *SITUACION_LABORAL* classes on task7b.

Team	F_1	P	R	System Summary
16	**0.87**	**0.8701**	**0.87**	BERTweet with fast gradient method
18	0.85	**0.8724**	0.8214	BERT-Large, BlueBERT with adversarial fine-tuning
27	0.84	**0.8706**	0.8058	BioBERT with data augmentation
-	0.85	0.86	0.837	-

Table 8: Evaluation results for Task 8: Classification of self-reported breast cancer posts on Twitter. Metrics show F_1-scores (F_1), precision (P), and recall (R) for the *self-reports* class.

ment in F_1-score.

4 Conclusion

This paper presents an overview of the sixth SMM4H shared tasks held in 2021. The shared tasks hosted a total of eight tasks with 12 individual tasks in total. With 40 teams participating in the shared tasks, we find that interest in tasks grew by 38% from the previous year. Analyzing the methods in the submitted systems, we find that the best systems used transformer based models such as BERT and RoBERTa with various techniques for addressing class imbalance. Details of individual systems are available as system description papers cited in Appendix A.

Acknowledgements

The work for #SMM4H 2020 at the University of Pennsylvania was supported by the National Institutes of Health (NIH) National Library of Medicine (NLM) [grant number R01LM011176]. The work on the Russian set of tweets was done at Kazan Federal University and supported by the Russian Science Foundation [grant number 18-11-00284]. The authors would also like to thank Alexis Upshur for her contribution to annotating tweets, Dmitry Ustalov and other members of the *Yandex.Toloka* team for providing credits for the crowd-sourced annotation of Russian tweets, and all those who reviewed system description papers. The work at the Barcelona Supercomputing Center was supported by the Encargo of the Spanish National Plan for the Advancement of Language Technology (PlanTL).

References

Alham Fikri Aji, Made Nindyatama Nityasya, Haryo Akbarianto Wibowo, Radityo Eko Prasojo, and Tirana Fatyanosa. 2021. BERT Goes Brrr: A Venture Towards the Lesser Error in Classifying Medical Self-Reporters on Twitter. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 58–64.

Juan M. Banda, Gurdas Viguruji Singh, Osaid H. Alser, and Daniel Prieto-Alhambra. 2020a. Long-term patient-reported symptoms of covid-19: an analysis of social media data. *medRxiv*.

Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Katya Artemova, Elena Tutubalina, and Gerardo Chowell. 2020b. A large-scale covid-19 twitter chatter dataset for open scientific research – an international collaboration.

Pavel Blinov. 2021. Text Augmentation Techniques in Drug Adverse Effect Detection Task. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 0–0.

Sergio Santamaría Carrasco and Roberto Cuervo Rosillo. 2021. Word Embeddings, Cosine Similarity and Deep Learning for Identification of Professions & Occupations in Health-related Social Media. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 74–76.

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*.

Joseph Cornelius, Tilia Ellendorff, and Fabio Rinaldi. 2021. Approaching SMM4H with auto-regressive language models and back-translation. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 146–148.

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.

Frances Adriana Laureano De Leon, Harish Tayyar Madabushi, and Mark Lee. 2021. UoB at ProfNER 2021: Data Augmentation for Classification Using Machine Translation. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 115–117.

George-Andrei Dima, Dumitru-Clementin Cercel, and Mihai Dascalu. 2021. Transformer-based Multi-Task Learning for Adverse Effect Mention Analysis in Tweets. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 44–51.

Mohab Elkaref and Lamiece Hassan. 2021. A Joint Training Approach to Tweet Classification and Adverse Effect Extraction and Normalization for SMM4H 2021. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 91–94.

David Carreto Fidalgo, Daniel Vila-Suero, Francisco Aranda Montes, and Ignacio Talavera Cepeda. 2021. System description for ProfNER - SMM4H: Optimized finetuning of a pretrained transformer and word vectors. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 69–73.

Max Fleming, Priyanka Dondeti, Caitlin Dreisbach, and Adam Poliak. 2021. Fine-tuning Transformers for Identifying Self-Reporting Potential Cases and Symptoms of COVID-19 in Tweets. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 131–134.

Yuting Guo, Yao Ge, Mohammed Ali Al-Garadi, and Abeed Sarker. 2021. Pre-trained Transformer-based Classification and Span Detection Models for Social Media Health Applications. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 52–57.

Andrey Gusev, Anna Kuznetsova, Anna Polyanskaya, and Egor Yatsishin. 2020. Bert implementation for detecting adverse drug effects mentions in russian. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 46–50.

Zongcheng Ji, Tian Xia, and Mei Han. 2021. PAII-NLP at SMM4H 2021: Joint Extraction and Normalization of Adverse Drug Effect Mentions in Tweets. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 126–127.

Tanay Kayastha, Pranjal Gupta, and Pushpak Bhattacharyya. 2021. BERT based Adverse Drug Effect Tweet Classification. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 88–90.

Ari Z. Klein, Ilseyar Alimova, Ivan Flores, Arjun Magge, Zulfat Miftahutdinov, Anne-Lyse Minard, Karen O'connor, Abeed Sarker, Elena Tutubalina, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2020a. Overview of the fifth Social Media Mining for Health Applications (# SMM4H) shared tasks at COLING 2020. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 27–36.

Ari Z. Klein, Haitao Cai, Davy Weissenbacher, Lisa Levine, and Graciela Gonzalez-Hernandez. 2020b. A natural language processing pipeline to advance the use of Twitter data for digital epidemiology of adverse pregnancy outcomes. *Journal of Biomedical Informatics: X*, 8:100076.

Ari Z. Klein and Graciela Gonzalez-Hernandez. 2020. An annotated data set for identifying women reporting adverse pregnancy outcomes on Twitter. *Data in Brief*, 32:106249.

Ari Z. Klein, Arjun Magge, Karen O'Connor, Jesus Ivan Flores Amaro, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021. Towards using Twitter for tracking COVID-19: A natural language processing pipeline and exploratory data set. *Journal of Medical Internet Research*, 23(1):e25314.

Adarsh Kumar, Ojasv Kamal, and Susmita Mazumdar. 2021a. Adversities are all you need: Classification of self-reported breast cancer posts on Twitter using Adversarial Fine-tuning. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 112–114.

Deepak Kumar, Nalin Kumar, and Subhankar Mishra. 2021b. NLP@NISER: Classification of COVID19 tweets containing symptoms. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 102–104.

Lung-Hao Lee, Man-Chen Hung, Chien-Huan Lu, Chang-Hao Chen, Po-Lei Lee, and Kuo-Kai Shyu. 2021. Classification of Tweets Self-reporting Adverse Pregnancy Outcomes and Potential COVID-19 Cases Using RoBERTa Transformers. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 98–101.

Ying Luo, Lis Pereira, and Kobayashi Ichiro. 2021. OCHADAI at SMM4H-2021 Task 5: Classifying self-reporting tweets on potential cases of COVID-19 by ensembling pre-trained language models. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 123–125.

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*.

Zulfat Miftahutdinov, Andrey Sakhovskiy, and Elena Tutubalina. 2020. Kfu nlp team at smm4h 2020 tasks: Cross-lingual transfer learning with pretrained language models for drug reactions. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 51–56.

Anupam Mondal, Sainik Mahata, Monalisa Dey, and Dipankar Das. 2021. Classification of COVID19 tweets using Machine Learning Approaches. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 135–137.

Alberto Mesa Murgado, Ana Parras Portillo, Pilar Maite Martin López Úbeda, and Alfonso Urena-López. 2021. Identifying professions & occupations in Health-related Social Media using Natural Language Processing. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 141–145.

Atul Kr. Ojha, Priya Rani, Koustava Goswami, Bharathi Raja Chakravarthi, and John P. McCrae. 2021. ULD-NUIG at Social Media Mining for

Health Applications (#SMM4H) Shared Task 2021. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 149–152.

Victoria Pachón, Jacinto Mata Vázquez, and Juan Luís Domínguez Olmedo. 2021. Identification of profession & occupation in Health-related Social Media using tweets in Spanish. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 0–0.

Vasile Pais and Maria Mitrofan. 2021. Assessing multiple word embeddings for named entity recognition of professions and occupations in health-related social media. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 0–0.

Varad Pimpalkhute, Prajwal Nakhate, and Tausif Diwan. 2021. Transformers Models for Classification of Health-Related Tweets. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 118–122.

Sidharth Ramesh, Abhiraj Tiwari, Parthivi Choubey, Saisha Kashyap, Sahil Khose, Kumud Lakara, Nishesh Singh, and Ujjwal Verma. 2021. BERT based Transformers lead the way in Extraction of Health Information from Social Media. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 33–38.

Rajarshi Roychoudhury and Sudip Naskar. 2021. Fine-tuning BERT to classify COVID19 tweets containing symptoms. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 138–140.

Pedro Ruas, Vitor Andrade, and Francisco Couto. 2021. Lasige-BioTM at ProfNER: BiLSTM-CRF and contextual Spanish embeddings for Named Entity Recognition and Tweet Binary Classification. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 108–111.

Andrey Sakhovskiy, Zulfat Miftahutdinov, and Elena Tutubalina. 2021. KFU NLP Team at SMM4H 2021 Tasks: Cross-lingual and Cross-modal BERT-based Models for Adverse Drug Effects. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 39–43.

Felipe Soares, Marta Villegas, Aitor Gonzalez-Agirre, Martin Krallinger, and Jordi Armengol-Estapé. 2019. Medical word embeddings for spanish: Development and evaluation. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 124–133.

Alberto Valdes, Jesus Lopez, and Manuel Montes. 2021. UACH at SMM4H: a BERT based approach for classification of COVID-19 Twitter posts. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 65–68.

Davy Weissenbacher, Suyu Ge, Ari Klein, Karen O'Connor, Robert Gross, Sean Hennessy, and Graciela Gonzalez-Hernandez. 2020. Active neural networks to detect mentions of changes to medication treatment in social media. *medRxiv*.

Usama Yaseen and Stefan Langer. 2021. Neural Text Classification and Stacked Heterogeneous Embeddings for Named Entity Recognition in SMM4H 2021. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 83–87.

Tong Zhou, Zhucong Li, Zhen Gan, Baoli Zhang, Yubo Chen, Kun Niu, Jing Wan, Kang Liu, Jun Zhao, Yafei Shi, Weifeng Chong, and Shengping Liu. 2021. Classification, Extraction, and Normalization : CASIA_Unisound Team at the Social Media Mining for Health 2021 Shared Tasks. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 77–82.

Appendix A. Team Numbers and System Description Papers

Team	System Description Paper
1	(Pais and Mitrofan, 2021)
2	(Pachón et al., 2021)
3	(Blinov, 2021)
4	(Ramesh et al., 2021)
5	(Carrasco and Rosillo, 2021)
6	(Fleming et al., 2021)
7	(Mondal et al., 2021)
8	(Valdes et al., 2021)
9	(Roychoudhury and Naskar, 2021)
10	(Yaseen and Langer, 2021)
11	(Kumar et al., 2021b)
12	(Aji et al., 2021)
13	(Luo et al., 2021)
14	(Murgado et al., 2021)
15	(Lee et al., 2021)
16	(Zhou et al., 2021)
17	(Fidalgo et al., 2021)
18	(Kumar et al., 2021a)
19	(De Leon et al., 2021)
20	(Kayastha et al., 2021)
21	(Ruas et al., 2021)
22	(Guo et al., 2021)
23	(Sakhovskiy et al., 2021)
24	(Elkaref and Hassan, 2021)
25	(Cornelius et al., 2021)
26	(Dima et al., 2021)
27	(Pimpalkhute et al., 2021)
28	(Ji et al., 2021)
29	(Ojha et al., 2021)

Table 9: Key for identifying teams in the Results section. Identifier in the parenthesis is the publication id associated with the SMM4H Workshop.