Overview of the Fifth Social Media Mining for Health Applications (#SMM4H) Shared Tasks at COLING 2020

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

Ilseyar Alimova
Kazan Federal University
Kazan, Russia

Ivan Flores
University of Pennsylvania
Philadelphia, PA, USA

Arjun Magge
University of Pennsylvania
Philadelphia, PA, USA

Zulfat Miftahutdinov
Kazan Federal University
Kazan, Russia

Anne-Lyse Minard
University of Orléans, LLL-CNRS
Orléans, France

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

Abeed Sarker
Emory University
Atlanta, GA, USA

Elena Tutubalina
Kazan Federal University
Kazan, Russia

Davy Weissenbacher
University of Pennsylvania
Philadelphia, PA, USA

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

{ariklein, ivan.flores, arjun.magge, karoc, dweissen, gragon}@pennmedicine.upenn.edu
{alimovailseyar, zulfatmi, tutubalinaev}@gmail.com
anne-lyse.minard@univ-orleans.fr, abeed@dbmi.emory.edu

Abstract

The vast amount of data on social media presents significant opportunities and challenges for utilizing it as a resource for health informatics. The fifth iteration of the Social Media Mining for Health Applications (#SMM4H) shared tasks sought to advance the use of Twitter data (tweets) for pharmacovigilance, toxicovigilance, and epidemiology of birth defects. In addition to re-runs of three tasks, #SMM4H 2020 included new tasks for detecting adverse effects of medications in French and Russian tweets, characterizing chatter related to prescription medication abuse, and detecting self reports of birth defect pregnancy outcomes. The five tasks required methods for binary classification, multi-class classification, and named entity recognition (NER). With 29 teams and a total of 130 system submissions, participation in the #SMM4H shared tasks continues to grow.

1 Introduction

The aim of the Social Media Mining for Health Applications (#SMM4H) shared tasks is to take a community-driven approach to addressing natural language processing (NLP) challenges of utilizing social media data for health informatics, including informal, colloquial expressions of clinical concepts, noise, data sparsity, ambiguity, and multilingual posts. The fifth iteration of the #SMM4H shared tasks consisted of five tasks involving mining health-related information from Twitter data (tweets): automatic classification of tweets that mention medications (Task 1), automatic classification of multilingual tweets that report adverse effects of a medication (Task 2), with sub-tasks for distinct sets of tweets posted in English (Task 2a), French (Task 2b), and Russian (Task 2c), automatic extraction and normalization of adverse effects in English tweets (Task 3), automatic characterization of chatter related to prescription medication abuse in tweets (Task 4), and automatic classification of tweets self-reporting a birth defect pregnancy outcome (Task 5).

Teams could register for one or multiple tasks. In total, 57 teams registered for at least one task. To develop their systems, teams were provided with annotated training and validation sets of tweets for

Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 27–36
Barcelona, Spain (Online), December 12, 2020.

each task. For the final evaluation, teams were provided with an unlabeled test set for each task, and were allowed up to four days to submit the predictions of their systems to CodaLab[1]—a platform that facilitates data science competitions. Each team was allowed to submit up to three sets of predictions per task. In total, 29 of the 57 registered teams submitted at least one set of predictions. More specifically, 16 teams participated in Task 1 (40 submissions), 17 teams in Task 2a (35 submissions), 5 teams in Task 2b (7 submissions), 7 teams in Task 2c (14 submissions), 7 teams in Task 3 (15 submissions), 3 teams in Task 4 (9 submissions), and 4 teams in Task 5 (10 submissions). In Section 2, we will briefly describe the tasks. In Section 3, we will present the performance and a brief summary of each team's best-performing system for each task. Appendix A provides the system description papers corresponding to the team numbers used in Section 3.

2 Tasks

2.1 Task 1: Automatic Classification of Tweets that Mention Medications

Task 1 is a binary classification task that involves distinguishing tweets that mention a medication or dietary supplement (annotated as "1") from those that do not (annotated as "0"). For this task, we used the definition of drug products and dietary supplements provided by the FDA (U.S. Food and Drug Administration, 2017). For the #SMM4H 2018 shared tasks (Weissenbacher et al., 2018), a data set was used that contained an artificially balanced distribution of the two classes. For #SMM4H 2020, the data set represents their natural, highly imbalanced distribution (Weissenbacher et al., 2019). Evaluating classifiers on this data set models more closely the detection of tweets that mention medications in practice. The training set contains 69,272 tweets, with only 181 (0.3%) tweets that mention a medication. The 9622 training tweets from #SMM4H 2018 were also provided, with 4975 tweets that mention a medication. The test set contains 29,687 tweets, with only 77 (0.3%) tweets that mention a medication. Systems were evaluated based on the F_1-score for the "positive" class (i.e., tweets that mention a medication).

2.2 Task 2: Automatic Classification of Multilingual Tweets that Report Adverse Effects

Task 2 is a binary classification task that involves distinguishing tweets that report an adverse effect of a medication (annotated as "1") from those that do not (annotated as "0"), with three sub-tasks for distinct sets of tweets posted in English, French, and Russian. The training set for the long-running, English-language version of this #SMM4H shared task contains 25,678 tweets, with 2377 (9.3%) tweets that report an adverse effect of a medication. The test set contains 4759 tweets, with 194 (4.1%) tweets that report an adverse effect.

For the French sub-task, the training set contains 2426 tweets, with only 39 (1.6%) tweets that report an adverse effect. The test set contains 607 tweets, with only 10 (1.6%) tweets that report an adverse effect. Inter-annotator agreement, based on dual annotations of 848 tweets by three annotators, was 0.61 and 0.69, for each of the two pairs of annotators.

For the Russian sub-task, the training set contains 7612 tweets, with 666 (8.7%) tweets that report an adverse effect. The test set contains 1903 tweets, with 166 (8.7%) tweets that report an adverse effect. 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. Inter-annotator agreement was 0.49 (Cohen's kappa). Systems were evaluated based on the F_1-score for the "positive" class (i.e., tweets that report an adverse effect).

2.3 Task 3: Automatic Extraction and Normalization of Adverse Effects in English Tweets

Task 3 is a named entity recognition (NER) and entity normalization task that involves detecting the span of text within a tweet that reports an adverse effect of a medication, and normalizing the adverse effect to a unique Medical Dictionary for Regulatory Activities (MedDRA)[3] version 21.1 preferred term (PT) ID. The training set contains 2806 tweets, with 1829 (65%) tweets that report an adverse effect (annotated

[1] https://codalab.org/
[2] https://toloka.yandex.ru/
[3] https://www.meddra.org/

as "ADR"). For each tweet in the training set that reports an adverse effect, the span of text containing the adverse effect, the character offsets of that span of text, and the MedDRA ID of the adverse effect. The test set contains 1156 tweets, with 970 (84%) that report an adverse effect. Systems were evaluated based on their F_1-score, where a true positive is both the correct adverse effect (either partially or exactly matching the actual character offsets) and the correct MedDRA ID.

2.4 Task 4: Automatic Characterization of Prescription Medication Abuse Chatter in Tweets

Task 4 is a multi-class classification task that involves automatically distinguishing tweets mentioning potentially abuse-prone medications into one of four categories: (1) potential abuse/misuse (annotated as "A"), (2) non-abuse/misuse consumption (annotated as "C"), (3) medication mention only without any indication of consumption (annotated as "M"), and (4) unrelated (annotated as "U"). The medications mentioned in the tweets include prescription opioids, benzodiazepines, atypical anti-psychotics, central nervous system stimulants, and GABA (gamma aminobutyric acid) analogues. The training set contains 13,172 tweets: (1) 2133 (16%) "A" tweets, (2) 3668 (28%) "C" tweets, (3) 6843 (52%) "M" tweets, and (4) 528 (4%) "U" tweets. The test set contains 3271 tweets: (1) 503 (15%) "A" tweets, (2) 919 (28%) "C" tweets, (3) 1722 "M" (53%) tweets, and (4) 127 (4%) "U" tweets. Additional details about the data set, including the annotation process, annotation guidelines, and inter-annotator agreements, are presented in recent work (O'Connor et al., 2020). Systems were evaluated based on the F_1-score for the "potential misuse/abuse" ("A") class.

2.5 Task 5: Automatic Classification of Tweets Reporting a Birth Defect Pregnancy Outcome

Task 5 is a multi-class classification task that involves automatically distinguishing three classes of tweets that mention birth defects (Klein et al., 2018): (1) "defect" tweets refer to the user's child and indicate that he or she has the birth defect mentioned in the tweet (annotated as "1"); (2) "possible defect" tweets are ambiguous about whether someone is the user's child and/or has the birth defect mentioned in the tweet (annotated as "2"); (3) "non-defect" tweets merely mention birth defects (annotated as "3"). The training set contains 18,397 tweets: 966 (5%) "defect" tweets, 1041 (6%) "possible defect" tweets, and 16,390 (89%) "non-defect" tweets. The test set contains 4602 tweets: 244 (5%) "defect" tweets, 258 (6%) "possible defect" tweets, and 4100 (89%) "non-defect" tweets. Inter-annotator agreement, based on dual annotations for 21,727 of the tweets, was 0.86 (Cohen's kappa). Systems were evaluated based on the micro-averaged F_1-score for the "defect" and "possible defect" classes.

3 Results

3.1 Task 1: Automatic Classification of Tweets that Mention Medications

Table 1 presents the precision, recall, and F_1-score for the "positive" class (i.e., tweets that mention a medication), for each of the 16 team's best-performing system for Task 1. The majority of teams used a transformer-based architecture. Among these teams, the difference in performance seems to be based on the corpora used to pre-train the transformers, and the strategies used the address the high degree of class imbalance. The results suggest that imbalanced data remains a challenge for training deep neural network classifiers. The best-performing system for this task in #SMM4H 2018 (Weissenbacher et al., 2018) achieved an F_1-score of 0.918 (Chuhan et al., 2018) using an artificially balanced data set, while the best-performing system in #SMM4H 2020 achieved an F_1-score of 0.854. Nonetheless, advances in transformer-based architectures and strategies for addressing class imbalance have improved upon the baseline F_1-score of 0.788 (Weissenbacher et al., 2019).

3.2 Task 2: Automatic Classification of Multilingual Tweets that Report Adverse Effects

3.2.1 Automatic Classification of English Tweets that Report Adverse Effects

Table 2 presents the precision, recall, and F_1-score for the "positive" class (i.e., English tweets that report an adverse effect of a medication), for each of the 17 team's best-performing system for Task 2a. As in Task 1, the majority of teams used a transformer-based architecture. In particular, most of the better-

Team	F_1	P	R	System Summary
8	**0.85**	0.84	0.87	BERT, Bio+Clinical BERT, sub-corpus ensemble, SMM4H'18 corpus
3	0.80	0.77	0.83	RoBERTa pre-trained on tweets, ensemble, SMM4H'18 corpus
21	0.80	0.80	0.79	RoBERTa pre-trained on tweets
2	0.77	0.71	0.83	BERT, SMM4H'18 corpus
14	0.76	0.73	0.79	ELECTRA, decision tree, data augmentation, ensemble, SMM4H'18 corpus
17	0.76	0.82	0.70	BERT, DrugBank
18	0.76	0.77	0.74	RoBERTa pre-trained on biomedical literature
12	0.74	0.66	0.83	RoBERTa pre-trained on biomedical literature, over-sampling, SMM4H'18 corpus
23	0.72	0.84	0.62	BERT, BiLSTM, SMM4H'18 corpus
19	0.71	0.79	0.64	BioBERT pre-trained on tweets, sub-corpus ensemble, class weights, SMM4H'18 corpus
28	0.66	0.75	0.58	NA
9	0.64	0.74	0.56	decision tree, word and character 15-grams
29	0.60	0.57	0.64	NA
6	0.56	**0.86**	0.42	BioBERT, data augmentation, ensemble
22	0.45	0.57	0.38	NA
16	0.05	0.02	**0.90**	SVM, sent2vec sentence and bi-gram embeddings pre-trained on tweets, under-sampling

Table 1: Task 1 system summaries and F_1-scores (F_1), precision (P), and recall (R) for the "positive" class (i.e., tweets mentioning medications).

Team	F_1	P	R	System Summary
21	**0.64**	0.62	0.65	RoBERTa
25	0.58	**0.63**	0.54	EnDR-BERT, ensemble
10	0.58	0.52	0.65	RoBERTa, SMM4H'17 and SMM4H'19 corpora
5	0.57	0.50	0.66	RoBERTa
4	0.56	0.50	0.63	RoBERTa
7	0.56	0.56	0.55	RoBERTa, sub-corpus ensemble, rules
17	0.55	0.47	0.65	BERT, DrugBank, MedlinePlus, TransE MeSH representations
6	0.54	0.49	0.60	BioBERT, data augmentation, ensemble
1	0.51	0.48	0.54	CLAPA, BERT
22	0.48	0.44	0.53	SBERT RoBERTa sentence embeddings, class weights
2	0.47	0.58	0.40	BERT, SMM4H'20 Task 3 corpus
19	0.37	0.26	0.60	BioBERT pre-trained on tweets
20	0.35	0.28	0.46	CNN, GloVe word embeddings pre-trained on tweets, under-sampling
16	0.32	0.19	**0.87**	SVM, sent2vec sentence and bi-gram embeddings pre-trained on tweets, under-sampling
28	0.31	0.23	0.51	NA
15	0.31	0.31	0.31	logistic regression, feature engineering
29	0.27	0.16	0.79	NA

Table 2: Task 2a (English) system summaries and F_1-scores (F_1), precision (P), and recall (R) for the "positive" class (i.e., tweets reporting an adverse effect of a medication).

performing systems used RoBERTa-based models (Liu et al., 2019), with the best-performing system achieving an F_1-score of 0.64.

3.2.2 Automatic Classification of French Tweets that Report Adverse Effects

Table 3 presents the precision, recall, and F_1-score for the "positive" class (i.e., French tweets that report an adverse effect of a medication), for each of the five team's best-performing system for Task 2b. The highest F_1-score for the French-language version of this task is considerably lower than the highest F_1-scores for the automatic classification of adverse effects in English (0.64) and Russian (0.51) tweets. The difficulty of this task is further underscored by the fact that two teams were not able to detect any tweets reporting an adverse effect. This difficulty may be due to the small size of the training data and the high degree of class imbalance. To address the imbalanced data, Team 22 used a Bayesian optimization approach to class weighting, and Team 16 used under-sampling of the majority class.

3.2.3 Automatic Classification of Russian Tweets that Report Adverse Effects

Table 4 presents the precision, recall, and F_1-score for the "positive" class (i.e., Russian tweets that report an adverse effect of a medication), for each of the seven team's best-performing system for Task 2c. Teams 26 and 25 achieve the highest F_1-scores (0.51). Both teams used ensembles of BERT-based Russian language models from the DeepPavlov library (Burtsev et al., 2018). In addition, both teams used manually annotated drug reviews from the RuDREC corpus (Tutubalina et al., 2020) as additional

Team	F_1	P	R	System Summary
22	**0.17**	0.15	0.20	SBERT DistilBERT sentence embeddings, class weights
15	0.15	**0.33**	0.10	logistic regression, feature engineering
16	0.07	0.04	**0.60**	tree-based ensemble, LASER sentence embeddings, under-sampling
4	0.00	0.00	0.00	camemBERT
29	0.00	0.00	0.00	NA

Table 3: Task 2b (French) system summaries and F_1-scores (F_1), precision (P), and recall (R) for the "positive" class (i.e., tweets reporting an adverse effect of a medication).

training data, and Team 25 also used English drug reviews from the PsyTAR corpus ((Zolnoori et al., 2019).

Team	F_1	P	R	System Summary
26	**0.51**	0.45	0.60	Conversational RuBERT, under-sampling, ensemble, RuDReC corpus
25	0.51	**0.54**	0.48	EnRuDR-BERT, ensemble, bilingual training, RuDReC and PsyTAR corpora
5	0.48	0.36	0.70	RuBERT
22	0.42	0.35	0.55	SBERT DistilBERT sentence embeddings, class weights
4	0.36	0.34	0.40	RuBERT
28	0.36	0.29	0.46	NA
16	0.35	0.22	**0.89**	SVM, LASER sentence embeddings, under-sampling

Table 4: Task 2c (Russian) system summaries and F_1-scores (F_1), precision (P), and recall (R) for the "positive" class (i.e., tweets reporting an adverse effect of a medication).

3.3 Task 3: Automatic Extraction and Normalization of Adverse Effects in English Tweets

Table 5 presents the F_1-scores for the NER-based extraction of adverse effect text spans, and the precision, recall, and F_1-scores for the normalization to the MedDRA ID, for each of the seven team's best-performing systems for Task 3. Team 25 outperformed the other teams for all the presented performance metrics. For the NER-based extraction, they used a transformer-based architecture with domain-specific models, dictionary-based features, and additional training data from the CADEC corpus (Karimi et al., 2015). For normalization, they used a domain-specific, BERT-based classifier, additional training data, and similarity metrics comparing BERT-based word embeddings of Unified Medical Language System (UMLS) concepts and extracted NERs. Several other teams used similar approaches, so the performance of Team 26 might be attributed to their language models that were pre-trained specifically for detecting adverse drug reactions.

Team	E F_1	N F_1	N P	N R	System Summary
25	**0.76**	**0.46**	**0.48**	**0.45**	EnDR-BERT, dictionary, BERT-based similarity metrics, CADEC
2	0.73	0.38	0.34	0.44	BERT, CADEC, SMM4H'17 corpus
10	0.69	0.35	0.33	0.38	RoBERTa, multi-task learning
4	0.58	0.22	0.24	0.20	SciBERT/BioBERT/BERT ensemble, fastText-based similarity metrics, CADEC
1	0.46	0.20	0.35	0.14	BiLSTM, CRF, GloVe and EXT word embeddings, QuickUMLS
27	0.56	0.15	0.15	0.14	NA
16	0.16	0.00	0.00	0.00	dictionary

Table 5: Task 3 system summaries, F_1-scores (F_1) for adverse effect extraction (E), and F_1-scores (F_1), precision (P), and recall (R) for adverse effect normalization (N).

3.4 Task 4: Automatic Characterization of Prescription Medication Abuse Chatter in Tweets

Table 6 presents the precision, recall, and F_1-scores for the "potential abuse/misuse" class, for each team's best-performing system for Task 4. Team 13 achieved the highest F_1-score (0.51) using a CNN, fastText word embeddings, and data augmentation by means of manufacturing tweets that are semantically similar to the training data. This F_1-score, however, is lower than the F_1-score (0.67) of a stacked ensemble of BERT (Devlin et al., 2019), ALBERT (Lan et al., 2020), and RoBERTa models, presented in recent work (Ali Al-Garadi et al., 2020).

Team	F_1	P	R	System Summary
13	**0.51**	**0.53**	0.50	CNN, fastText word embeddings, data augmentation
1	0.49	0.46	0.51	SVM, under-sampling
16	0.46	0.35	**0.67**	SVM, sent2vec sentence and bi-gram embeddings pre-trained on tweets, under-sampling

Table 6: Task 4 system summaries and F_1-scores (F_1), precision (P), and recall (R) for the "potential abuse/misuse" class.

3.5 Task 5: Automatic Classification of Tweets Reporting a Birth Defect Pregnancy Outcome

Table 7 presents the micro-averaged precision, recall, and F_1-score for the "defect" and "possible defect" classes, for each team's best-performing system. Teams 6 and 24 achieved the highest micro-averaged F_1-scores (0.69). While Team 6 achieved a higher micro-averaged recall (0.73) than Team 24 (0.67) using a hard-voting ensemble of nine BioBERT-based models, Team 24 achieved a higher micro-averaged precision (0.71) than Team 6 (0.65) using ELMo word embeddings and data-specific resources for modeling birth defects, pregnancy-related information, people's names, and family relations. Team 19 also achieved a higher micro-averaged recall (0.69) than Team 24 (0.67) using BioBERT, but achieved a substantially lower micro-averaged precision (0.56) than Team 24 (0.71). Overall, for this imbalanced data, models based on contextualized word representations—BioBERT (Lee et al., 2020a) or ELMo (Peters et al., 2018)—outperformed a CNN-BiGRU neural network with GloVe word embeddings (Pennington et al., 2014). Recent work (Klein et al., 2019) presents baseline F_1-scores of an SVM classifier for the "defect" (0.65) and "possible defect" (0.51) classes.

Team	F_1	P	R	System Summary
6	**0.69**	0.65	**0.73**	BioBERT, data augmentation, ensemble
24	0.69	**0.71**	0.67	ELMo, GCNN, ANNIE NER, medical and family relations lexicons
19	0.62	0.56	0.69	BioBERT pre-trained on tweets
11	0.58	0.54	0.64	GloVe word and hashtag embeddings pre-trained on tweets, CNN, BiGRU

Table 7: Task 5 system summaries and micro-averaged F_1-score (F_1), precision (P), and recall (R) for the "defect" and "possible defect" classes.

4 Conclusion

This paper presented an overview of the #SMM4H 2020 shared tasks. With 29 teams and a total of 130 system submissions, participation in the #SMM4H shared tasks continues to grow. All of the teams with the best-performing system for each task used deep learning-based systems, most of which were transformer-based architectures. The system description papers that are cited in Appendix A were each peer-reviewed by two reviewers and provide further details about 26 teams' systems.

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 at Kazan Federal University was 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 crowdsourced annotation of Russian tweets, and all those who reviewed system description papers.

References

Olanrewaju Tahir Aduragba, Jialin Yu, Gautham Senthilnathan, and Alexandra Cristea. 2020. Sentence contextual encoder with BERT and BiLSTM for automatic classification with imbalanced medication tweets. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 165–167.

Mohammed Ali Al-Garadi, Yuan-Chi Yang, Haitao Cai, Yucheng Ruan, Karen O'Connor, Graciela Gonzalez-Hernandez, Jeanmarie Perrone, and Abeed Sarker. 2020. Text classification models for the automatic detection of nonmedical prescription medication use from social media. *medRxiv*.

Yandrapati Prakash Babu and Rajagopal Eswari. 2020. Identification of medication tweets using domain-specific pre-trained language models. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 128–130.

Parsa Bagherzadeh and Sabine Bergler. 2020. CLaC at SMM4H 2020: Birth defect mention detection. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 168–170.

Yang Bai and Xiaobing Zhou. 2020. Automatic detecting for health-related Twitter data with BioBERT. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 63–69.

Pavel Blinov and Manvel Avetisian. 2020. Transformer models for drug adverse effects detection from tweets. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 110–112.

Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yurii Kuratov, Denis Kuznetsov, et al. 2018. Deeppavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127.

Silvia Casola and Alberto Lavelli. 2020. FBK@SMM4H2020: RoBERTa for detecting medications on Twitter. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 101–103.

Wu Chuhan, Wu Fangzhao, Liu Junxin, Wu Sixing, Huang Yongfeng, and Xie Xing. 2018. Detecting tweets mentioning drug name and adverse drug reaction with hierarchical tweet representation and multi-head self-attention. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop and Shared Task*, pages 34–37.

Huong N. Dang, Kahyun Lee, Sam Henry, and Özlem Uzuner. 2020. Ensemble BERT for classifying medication-mentioning tweets. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 37–41.

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. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.

George-Andrei Dima, Andrei-Marius Avram, and Dumitru-Clementin Cercel. 2020. Approaching SMM4H 2020 with ensembles of BERT flavours. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 153–157.

Lucie Gattepaille. 2020. How far can we go with just out-of-the-box BERT models? In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 95–100.

Oguzhan Gencoglu. 2020. Sentence transformers and Bayesian optimization for adverse drug effect detection from Twitter. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 161–164.

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 (#SMM4H) Workshop & Shared Task*, pages 46–50.

Katikapalli Subramanyam Kalyan and Sivanesan Sangeetha. 2020. Want to identify, extract and normalize adverse drug reactions in tweets? Use RoBERTa. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 121–124.

Sarvnaz Karimi, Alejanrdo Metke-Himenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug effect annotations. *Journal of Biomedical Informatics*, 55:73–81.

Sedigheh Khademi, Pari Delir Haghighi, and Frada Burstein. 2020. Adverse drug reaction detection in Twitter using RoBERTa and rules. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 113–117.

Ari Z. Klein, Abeed Sarker, Haitao Cai, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2018. Social media mining for birth defects research: A rule-based, bootstrapping approach to collecting data for rare health-related events on Twitter. *Journal of Biomedical Informatics*, 87:68–78.

Ari Z. Klein, Abeed Sarker, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2019. Towards scaling Twitter data for digital epidemiology of birth defects. *npj Digital Medicine*, 2:1–9.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Jinhyuk Lee, Wonjin Yoon, Sundong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020a. BioBERT: A pretrained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Lung-Hao Lee, Po-Han Chen, Hao-Chuan Kao, Ting-Chun Hung, Po-Lei Lee, and Kuo-Kai Shyu. 2020b. Medication mention detection in tweets using ELECTRA transformers and decision trees. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 131–133.

Mohamed Lichouri and Mourad Abbas. 2020. SpeechTrans@SMM4H'20: Impact of preprocessing and n-grams on automatic classification of tweets that mention medications. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 118–120.

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

Farhana Ferdousi Liza. 2020. Sentence classification with imbalanced data for health applications. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 138–145.

Darshini Mahendran, Cora Lewis, and Bridget T. McInnes. 2020. NLP@VCU: Identifying adverse effects in English tweets for unbalanced data. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 158–160.

Laiba Mehnaz. 2020. Automatic classification of tweets mentioning a medication using pre-trained sentence encoders. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 150–152.

Isabel Metzger, Emir Y. Haskovic, Allison Black, Whitley M. Yi, Rajat S. Chandra1, Mark T. Rutledge, William McMahon, and Yindalon Aphinyanaphongs. 2020. SMM4H Shared Task 2020 - A hybrid pipeline for identifying prescription drug abuse from Twitter: Machine learning, deep learning, and post-processing. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 57–62.

Zulfat Miftahutdinova, 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 (#SMM4H) Workshop & Shared Task*, pages 51–56.

Karen O'Connor, Abeed Sarker, Jeanmarie Perrone, and Graciela Gonzalez Hernandez. 2020. Promoting reproducible research for characterizing nonmedical use of medications through data annotation: Description of a Twitter corpus and guidelines. *Journal of Medical Internet Research*, 22(2):e15861.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, and Matt Gardner. 2018. Deep contextualized word representations. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2227–2237.

Saichethan Miriyala Reddy. 2020. Detecting tweets reporting birth defect pregnancy outcome using two-view CNN RNN based architecture. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 125–127.

Sougata Saha, Souvik Das, Prashi Khurana, and Rohini K. Srihari. 2020. Autobots ensemble: Identifying and extracting adverse drug reaction from tweets using transformer based pipelines. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 104–109.

Ludovic Tanguy, Lydia-Mai Ho-Dac, Cécile Fabre, Roxane Bois, Touati Mohamed Yacine Haddad, Claire Ibarboure, Marie Joyau, François Le moal, Jade Moillic, Laura Roudaut, Mathilde Simounet, Irena Stankovic, and Mickaela Vandewaetere. 2020. LITL at SMM4H: An old-school feature-based classifier for identifying adverse effects in tweets. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 134–137.

Elena Tutubalina, Ilseyar Alimova, Zulfat Miftahutdinov, Andrey Sakhovskiy, Valentin Malykh, and Sergey Nikolenko. 2020. The Russian Drug Reaction Corpus and neural models for drug reactions and effectiveness detection in user reviews. *Bioinformatics*.

U.S. Food and Drug Administration. 2017. Drugs@fda glossary of terms. `https://www.fda.gov/drugs/drug-approvals-and-databases/drugsfda-glossary-terms`. [Drug; Drug Product; online, accessed 21-July-2020].

V.G.Vinod Vydiswaran, Deahan Yu, Xinyan Zhao, Ermioni Carr, Jonathan Martindale, Jingcheng Xiao, Noha Ghannam, Matteo Althoen, Alexis Castellanos, Neel Patel, and Daniel Vasquez. 2020. Identifying medication abuse and adverse effects from tweets: University of Michigan at #SMM4H 2020. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 90–94.

Chen-Kai Wang, You-Chen Zhang, Bo-Chun Xu, Bo-Hong Wang, You-Ning Xu, Po-Hao Chen, Hong-Jie Dai, and Chung-Hong Lee. 2020. ISLab system for SMM4H Shared Task 2020. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 42–45.

Davy Weissenbacher, Abeed Sarker, Michael J. Paul, and Graciela Gonzalez-Hernandez. 2018. Overview of the third social media mining for health (SMM4H) shared tasks at EMNLP 2018. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 13–16.

Davy Weissenbacher, Abeed Sarker, Ari Klein, Karen O'Connor, Arjun Magge, and Graciela Gonzalez-Hernandez. 2019. Deep neural networks for ensemble for detecting medication mentions in tweets. *Journal of the American Medical Informatics Association*, 26(12):1618–1626.

Xiaoyu Zhao, Ying Xiong, and Buzhou Tang. 2020. HITSZ-ICRC: A report for SMM4H shared task 2020-Automatic classification of medications and adverse effect in tweets . In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 146–149.

Maryam Zolnoori, Kin Wah Fung, Timothy B. Patrick, Paul Fontelo, Hadi Kharrazi, Anthony Faiola, Yi Shuan Shirley Wu, Christina E. Eldredge, Jake Luo, Mike Conway, et al. 2019. A systematic approach for developing a corpus of patient reported adverse drug events: A case study for SSRI and SNRI medications. *Journal of Biomedical informatics*, 90:103091.

Appendix A. Team Numbers and System Description Papers

Team	System Description Paper
1	(Vydiswaran et al., 2020)
2	(Gattepaille, 2020)
3	(Casola and Lavelli, 2020)
4	(Saha et al., 2020)
5	(Blinov and Avetisian, 2020)
6	(Bai and Zhou, 2020)
7	(Khademi et al., 2020)
8	(Dang et al., 2020)
9	(Lichouri and Abbas, 2020)
10	(Kalyan and Sangeetha, 2020)
11	(Reddy, 2020)
12	(Babu and Eswari, 2020)
13	(Metzger et al., 2020)
14	(Lee et al., 2020b)
15	(Tanguy et al., 2020)
16	(Liza, 2020)
17	(Zhao et al., 2020)
18	(Mehnaz, 2020)
19	(Dima et al., 2020)
20	(Mahendran et al., 2020)
21	(Wang et al., 2020)
22	(Gencoglu, 2020)
23	(Aduragba et al., 2020)
24	(Bagherzadeh and Bergler, 2020)
25	(Miftahutdinova et al., 2020)
26	(Gusev et al., 2020)
27	NA
28	NA
29	NA