SocBERT: A Pretrained Model for Social Media Text

Yuting Guo and Abeed Sarker
Department of Biomedical Informatics, School of Medicine
Emory University, Atlanta GA 30322, USA
yuting.guo@emory.edu
abeed@dbmi.emory.edu

Abstract

Pretrained language models (PLMs) on domain-specific data have been proven to be effective for in-domain natural language processing (NLP) tasks. Our work aimed to develop a language model which can be effective for the NLP tasks with the data from diverse social media platforms. We pretrained a language model on Twitter and Reddit posts in English consisting of 929M sequence blocks for 112K steps. We benchmarked our model and 3 transformer-based models—BERT, BERTweet, and RoBERTa on 40 social media text classification tasks. The results showed that although our model did not perform the best on all of the tasks, it outperformed the baseline model—BERT on most of the tasks, which illustrates the effectiveness of our model. Also, our work provides some insights of how to improve the efficiency of training PLMs.

1 Introduction

In recent years, pretraining language models (PLMs) such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have proven to be effective for a wide range of natural language processing (NLP) tasks. Domain adaptive pretraining (DAPT), also known as pretraining on domain-specific data, has been a commonly employed approach to enhancing model performance on tasks that are specific to a particular domain (Gururangan et al., 2020). Numerous efforts have been made to achieve this goal. For example, Lee et al. (2019) proposed BioBERT by pretraining BERT on a large biomedical corpus of PubMed abstracts, and demonstrated that it outperformed BERT on three representative biomedical text mining tasks. Alsentzer et al. (2019) attempted to adapt pretrained models for clinical text by training BioBERT on clinical notes, resulting in the creation of BioClinical_BERT (Leroy et al., 2017). Encouraged by the success of pretraining models in different domains, recent studies have developed pretrained models for social media NLP tasks. For example, Dai et al. (2020) built a model by further pretraining the model developed by Devlin et al. (2019) by further training the model on English tweets. Nguyen et al. (2020a) pretrained a transformer model named BERTweet by training the model on a large scale of English tweets from scratch. However, these models only involve Twitter data and may not be effective enough for social media data from other platforms such as Reddit and Facebook. To fill this gap, we trained a language model using both Twitter and Reddit data. We used 92GB text data including 20GB English tweets and 72GB Reddit comments. Our model was trained from scratch for 112K steps following the model architecture of RoBERTa-base. For evaluation, We benchmarked our model and 3 transformer-based models—BERT, BERTweet, and RoBERTa on 40 social media text classification tasks covering diverse health-related and non-health-related topics and from 6 social media platforms. The results showed that although our model did not perform the best on all of the tasks, it outperformed the baseline model—BERT on most of the tasks. It showed that pretraining on the in-domain data can benefit the model on the downstream tasks. To sum up, our contributions are as follows:

- We pretrained and released a transformer-based language model on Twitter and Reddit data which outperformed BERT on most of the benchmarking tasks.
- We benchmarked our model and 3 PLMs on 40 social media text classification tasks.
- We analyzed the influence of training time, data source, and task domains to different PLMs, which.
- Our work provided some insights of how to improve the efficiency of training PLMs.

The Fourth Workshop on Insights from Negative Results in NLP, pages 45–52

We call our final pretrained model SocBERT—an abbreviation of Social Media BERT.

2 Method

2.1 Data collection and preprocessing

We collected 92GB pre-training data including 20GB English tweets and 72GB English Reddit comments. The Twitter data were collected via Twitter streaming API and downloaded from the Achive team [1], and the Reddit comments were downloaded from Pushshift[2]. Because Reddit comments are usually longer than the maximum sequence limitation of the language model, we chunked the comments into sequence blocks, and each sequence block is limited to the maximum sequence length. In addition, we used the open source tool named preprocess-twitter (Paulus and Pennington) to preprocessing the data. The preprocessing includes lowercasing, normalization of numbers, usernames, urls, hashtags and text smileys, and adding extra marks for capital words, hashtags and repeated letters. We applied fastBPE (Sennrich et al., 2016) to tokenize the data and obtained a dictionary including 74K subwords which was used for the model pretraining.

2.2 Model architecture

We developed a masked language model (MLM) for pretraining and a classification model for benchmarking. MLM is an unsupervised task in which some of the tokens in a text sequence are randomly masked in the input and the objective of the model is to predict the masked text segments. The model architectures for the masked language model (MLM) and classification are the same as the work of Liu et al. (2019). Specifically, MLM consists of an encoder layer that embeds the text sequence as an embedding matrix consisting of token embeddings and an output layer with Softmax activation that predict the masked token based on the embeddings of the masked tokens. The classification model consists of the same encoder layer and an output layer with Softmax activation to predict classes based on the embedding of the [CLS] token.

[1] https://archive.org/details/twitterstream

[2] https://files.pushshift.io/reddit/comments/

3 Benchmarking Tasks

We utilized a total of 40 social media text classification tasks to establish a benchmark, which represents the most extensive collection of social media text classification tasks currently available to us. Manually annotated data for all these tasks were either publicly available or had been made available through shared tasks. The tasks covered diverse health-related and non-health-related topics including, but not limited to, adverse drug reactions (ADRs) (Sarker and Gonzalez, 2015a; Sarker et al., 2018b), cohort identification for breast cancer (Al-Garadi et al., 2020), non-medical prescription medication use (NPMU) (Al-Garadi et al., 2021), informative COVID-19 content detection (Nguyen et al., 2020b), medication consumption (Sarker et al., 2018a), pregnancy outcome detection (Klein and Gonzalez-Hernandez, 2020), symptom classification (Magge et al., 2021), suicidal ideation detection (Gaur et al., 2021), identification of drug addiction and recovery intervention (Ghosh et al., 2020b), signs of pathological gambling and self-harm detection (Parapar et al., 2021), sentiment analysis and factuality classification in e-health forums (Carrillo-de Albornoz et al., 2018), offensive language identification (Zampieri et al., 2019), cyberbullying detection (Kumar et al., 2018; Bhattacharya et al., 2020), sentiment analysis (Mohammad et al., 2018; Preoţiuc-Pietro et al., 2016), and sarcasm language detection (Ghosh et al., 2020a).

The full details including the source, evaluation metric, training and test set sizes, the number of classes, and the inter-annotator agreement (IAA) for each task, if available, are shown in Appendix A. Seventeen tasks involved binary classification, 13 involved three-class classification, and 10 involved four-, five-, six- or nine-class classification each. The datasets combined included a total of 252,655 manually-annotated instances, with 204,989 (80%) instances for training and 47,666 (20%) for evaluation. The datasets involved data from multiple social media platforms—22 from Twitter, 6 from MedHelp[3], 6 from Reddit, 3 from Facebook, 2 from Youtube, and 1 from WebMD[4]. For evaluation, we used the F_1-score of the positive class for binary classification and the micro-averaged F_1-score for other multi-class classification.

[3] https://www.medhelp.org/

[4] https://www.webmd.com/

4 Experiments

4.1 Language model settings

The language model training consists of two phases. At the first phase, we initialized the language model with random initialization and trained the model on 20GB English tweets and 54GB Reddit comments for 100K steps from scratch. However, During this process, we observed that it would be extremely time-consuming to train the model on the whole dataset using our computation resources, and we could not inspect the model during this process. Therefore, at the second phase, we changed our training strategy into splitting the data and sequentially training the model on a each split so that we could check the model after each round. [5] Specifically, we split the Reddit data into small datasets with 10M sequence blocks and then trained the model on each dataset for 10 epochs. At the time of publication of this work, we finished the training of 11 small datasets involving another 18GB Reddit data. The maximum sequence limitation of our model is 128, and the batch size is 8192. Other hyper-parameters were the same for the two phases, which followed the settings of RoBERTa-base (Liu et al., 2019). We refer to the checkpoint at the end of first phase as SocBERT-base and the checkpoint at the end of second phase as SocBERT-final. In summary, SocBERT-base was pretrained on 819M sequence blocks for 100K steps. SocBERT-final was pretrained on 929M (819M+110M) sequence blocks for 112K (100K+12K) steps.

4.2 Classification model settings

For classification, we performed a limited parameter search with the learning rate $\in \{2 \times 10^{-5}, 3 \times 10^{-5}\}$ and fine-tuned each model for 10 epochs. The rest of hyper-parameters were the same as Liu et al. (2019). Because initialization can have a significant impact on convergence in training deep neural networks, we ran each experiment three times with different random initializations. The model that achieved the median performance over the test set is reported. In addition, we experimented with BERT-base, BERTweet, and RoBERTa-base to better evaluate the effectiveness of our model.

[5] The first phase training took about two and half a month. At the second phase, each round of training took about one week. The GPU model we used was 32GB Tesla V100. We used 8 GPUs at the first phase and 1 GPU at the second phase because of the limited budget.

5 Results

5.1 Classification results

The full classification results are listed in 1. We treated BERT as the baseline model and compared other models with BERT. BERTweet achieved better results on 33 (83%) tasks, RoBERTa on 35 (88%) tasks, SocBERT-base on 30 (75%) tasks, and SocBERT-final on 31 (78%) tasks. Although slightly underperforming RoBERTa and BERTweet, both of SocBERT-base and SocBERT-final outperformed BERT. It showed that our pretraining model is effective on the classification tasks with social media data. The gap between our model and RoBERTa was predictable because RoBERTa was pretrained on a much larger data set (160GB), for longer time (500K steps) than our model, and the pretraining data of RoBERTa also covered the Reddit data in our dataset. Compared to BERTweet, which was pretrained on 160M sequence blocks for 950K steps, our model was pretrained on a larger set of data for shorter time. This suggests that the training time may have a higher impact than the training data size on large language model pretraining. In addition, we observed that SocBERT-final outperformed SocBERT-base on 20 tasks. Considering that the second phase contained only 12K steps, it is reasonable that the influence of the second phase of training was small. Although the strategy we used for the second phase of training allowed us to check the model without waiting for several months, future studies are required to assess whether the strategy of the second phase of training is as efficient as training the model on the whole dataset. Since SocBERT-base and SocBERT-final performed similarly, we performed analysis only on SocBERT-base later in this section.

5.2 Model Comparison

In order to explore the influence of the data source and task domain, we compared the model performance of SocBERT-base, BERTweet, and RoBERTa over the tasks from different social media platforms or focusing on different topics shown in Table 2. The results showed that SocBERT-base outperformed BERTweet on 13 tasks and outperformed RoBERTa on 9 tasks. Although SocBERT-base and RoBERTa underperformed BERTweet on most of the tasks from Twitter, SocBERT-base and RoBERTa performed better on most of the tasks from Reddit and MedHelp. This suggests classification performance is likely to improve if the

ID	Task	Source	BERT	BT	RB	Soc-b	Soc-f
1	ADR Detection (Sarker and Gonzalez, 2015b)	Twitter	59.6	64.7	62.2	60.1	**66.0**
2	Breast Cancer (Sarker et al., 2020)	Twitter	85.6	88.1	**88.6**	86.1	86.6
3	NPMU characterization (Ali Al-Garadi et al., 2020)	Twitter	57.2	**66.1**	61.3	64.2	61.2
4	WNUT-20-task2 (COVID-19 tweet detection) (Nguyen et al., 2020c)	Twitter	86.6	88.5	**88.8**	87.9	87.8
5	SMM4H-17-task1 (ADR detection) (Sarker et al., 2018b)	Twitter	45.4	51.4	**53.8**	51.0	50.2
6	SMM4H-17-task2 (medication consumption) (Sarker et al., 2018b)	Twitter	76.5	**79.8**	78.6	77.4	78.1
7	SMM4H-21-task1 (ADR detection) (Magge et al., 2021)	Twitter	**70.5**	65.6	69.2	63.1	63.1
8	SMM4H-21-task3a (regimen change on Twitter) (Magge et al., 2021)	Twitter	55.6	55.9	**57.9**	57.4	55.5
9	SMM4H-21-task3b (regimen change on WebMD) (Magge et al., 2021)	WebMD	86.8	**88.4**	88.2	87.9	87.8
10	SMM4H-21-task4 (adverse pregnancy outcomes) (Magge et al., 2021)	Twitter	86.8	88.9	**89.7**	86.8	88.2
11	SMM4H-21-task5 (COVID-19 potential case) (Magge et al., 2021)	Twitter	69.6	72.3	**76.5**	71.8	74.3
12	SMM4H-21-task6 (COVID-19 symptom) (Magge et al., 2021)	Twitter	97.6	98.4	**98.2**	97.8	97.8
13	SMM4H-22-task9 (self-reporting exact age) (Weissenbacher et al., 2022)	Reddit	94.0	93.4	**94.2**	91.5	93.3
14	Suicidal Ideation Detection (Gaur et al., 2021)	Reddit	71.7	73.0	78.0	76.7	**78.6**
15	Drug Addiction and Recovery Intervention (Ghosh et al., 2020b)	Reddit	73.3	75.4	77.0	75.9	**77.5**
16	eRisk-21-task1 (Signs of Pathological Gambling) (Parapar et al., 2021)	Reddit	82.7	85.1	85.4	86.1	**87.6**
17	eRisk-21-task2 (Signs of Self-Harm) (Parapar et al., 2021)	Reddit	76.7	78.5	78.9	77.3	78.9
18	Sentiment Analysis (Food Allergy) (Carrillo-de Albornoz et al., 2018)	MedHelp	77.0	**75.8**	**75.8**	73.9	73.9
19	Sentiment Analysis (Crohn'S Disease) (Carrillo-de Albornoz et al., 2018)	MedHelp	70.8	73.9	**78.3**	76.4	73.9
20	Sentiment Analysis (Breast Cancer) (Carrillo-de Albornoz et al., 2018)	MedHelp	63.5	63.3	64.2	61.8	**64.6**
21	Factuality Classification (Food Allergy) (Carrillo-de Albornoz et al., 2018)	MedHelp	69.9	72.0	73.7	72.7	**74.0**
22	Factuality Classification (Crohn'S Disease) (Carrillo-de Albornoz et al., 2018)	MedHelp	**77.6**	71.7	71.4	74.5	75.9
23	Factuality Classification(Breast Cancer) (Carrillo-de Albornoz et al., 2018)	MedHelp	43.8	45.5	**50.0**	46.9	49.2
24	OLID-1 (Zampieri et al., 2019)	Twitter	83.1	84.9	84.9	**85.5**	85.3
25	OLID-2 (Zampieri et al., 2019)	Twitter	56.7	**90.8**	89.2	90.0	89.6
26	OLID-3 (Zampieri et al., 2019)	Twitter	36.6	70.0	69.0	**70.9**	66.7
27	TRAC-1-1 (Kumar et al., 2018)	Facebook	58.1	**60.3**	56.8	59.6	56.9
28	TRAC-1-2 (Kumar et al., 2018)	Twitter	56.6	**65.4**	59.8	59.3	58.6
29	TRAC2-1 (Bhattacharya et al., 2020)	Youtube	73.6	74.7	**75.6**	73.3	75.1
30	TRAC2-2 (Bhattacharya et al., 2020)	Youtube	**86.6**	85.8	85.6	86.3	85.3
31	SemEval-2018 Task 1-4 (Mohammad et al., 2018)	Twitter	67.8	69.1	68.6	66.5	**74.8**
32	SemEval-2018 Task 1-2-1 (Mohammad et al., 2018)	Twitter	70.1	**76.3**	73.3	75.4	76.1
33	SemEval-2018 Task 1-2-2 (Mohammad et al., 2018)	Twitter	86.6	86.4	**87.1**	86.4	85.4
34	SemEval-2018 Task 1-2-3 (Mohammad et al., 2018)	Twitter	72.8	**79.0**	77.8	77.5	73.0
35	SemEval-2018 Task 1-2-4 (Mohammad et al., 2018)	Twitter	62.9	**70.3**	67.6	67.1	64.7
36	Valence CLS (Preoţiuc-Pietro et al., 2016)	Facebook	63.3	**71.1**	**71.1**	64.7	65.3
37	Arousal CLS (Preoţiuc-Pietro et al., 2016)	Facebook	65.6	**71.5**	69.6	65.8	69.9
38	Sarcasm-FigLang-Reddit (Ghosh et al., 2020a)	Reddit	62.3	**67.5**	66.1	63.6	65.6
39	Sarcasm-FigLang-Twitter (Ghosh et al., 2020a)	Twitter	76.2	77.6	**80.9**	79.8	75.4
40	Airline (sentiment analysis) (Crowdflower, 2016)	Twitter	85.1	**86.3**	85.8	85.4	85.3

Table 1: The results of BERT, BERTweet (BT), RoBERTa (RB), SocBERT-base (Soc-b), and SocBERT-final (Soc-f) on 40 classification tasks. The task details can be found in Appendix. The best result for each task is in bold.

pretraining of a model includes data from the same social media source as the downstream tasks.

	Total	Soc >BT	Soc >RB	RB >BT
All tasks	40	13	9	19
Social media platform				
Twitter	22	5	5	9
Reddit	6	3	1	5
MedHelp	6	4	1	4
Facebook	3	0	1	0
Youtube	2	1	1	1
WebMD	1	0	0	0
Task domain				
Health	23	8	3	16
Non-health	17	5	6	3

Table 2: The comparison of model performance of SocBERT-base (Soc), BERTweet (BT), and RoBERTa (RB) over the tasks from different social media platforms or focusing on different topics. The symbol $A > B$ denotes that the model A outperforms the model B.

Another interesting observation is that on the health-related tasks, RoBERTa largely outperformed SocBERT-base and BERTweet, and SocBERT-Tweet slightly outperformed BERTweet. The possible explanation is that the linguistic characteristics of the pretraining data of RoBERTa and SocBERT-base can be more diverse than BERTweet because BERTWeet used a single-source corpus for pretraining.

6 Discussion

Our work initially aimed to develop a PLM which can efficiently work for the data from different social media platforms. However, the results showed that our model could not perform the best on all of the tasks compared to BERTweet and RoBERTa. The possible reason was that the training time of our model was not sufficient because of the limited computing resources. It revealed the dilemma for small labs in academia to develop large language models which has been studied since large language models became popular in the NLP field (Xu, 2022). However, our work can provide some insights for the NLP studies about developing and applying PLMs. First, training the model on a relatively small dataset for longer time might be more efficient than training the model on a large set of data for shorter time. Second, pretraining the model on in-domain data may more efficiently improve the performance on downstream tasks than pretraining on out-of-domain data. Also, the language models pretrained on sufficiently large open-domain data can be effective on domain-specific tasks. We released our model SocBERT-base [6] and SocBERT-final [7] via Huggingface to help the NLP community conduct further studies in this field.

7 Conclusion

In this work, we pre-trained a transformer-based model from scratch on social media data and benchmarked the model on 40 text classification tasks with social media data. Although our model did not perform the best on all of the tasks, it outperformed the baseline model—BERT on most of the benchmarking tasks. It showed that our model can be efficient for the text classification tasks with social media data. It may be possible to further improve the model performance if we continue training the model more efficiently. Further work is required to improve the efficiency and reduce the cost of large language model training.

References

M.A. Al-Garadi, Y.-C. Yang, S. Lakamana, J. Lin, S. Li, A. Xie, W. Hogg-Bremer, M. Torres, I. Banerjee, and A. Sarker. 2020. *Automatic Breast Cancer Cohort Detection from Social Media for Studying Factors Affecting Patient-Centered Outcomes*, volume 12299 LNAI.

Mohammed Ali Al-Garadi, Yuan Chi Yang, Haitao Cai, Yucheng Ruan, Karen O'Connor, Gonzalez Hernandez Graciela, Jeanmarie Perrone, and Abeed Sarker. 2021. Text Classification Models for the Automatic Detection of Nonmedical Prescription Medication Use From Social Media. *BMC Medical Informatics and Decision Making*, 21(1):1–13.

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*.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh

[6] https://huggingface.co/sarkerlab/SocBERT-base

[7] https://huggingface.co/sarkerlab/SocBERT-final

Dawer, Bornini Lahiri, and Atul Kr. Ojha. 2020. Developing a Multilingual Annotated Corpus of Misogyny and Aggression. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 158–168, Marseille, France. European Language Resources Association (ELRA).

Jorge Carrillo-de Albornoz, Javier Rodriguez Vidal, and Laura Plaza. 2018. Feature Engineering for Sentiment Analysis in E-health Forums. *PLoS ONE*, 13(11):e0207996.

Crowdflower. 2016. Twitter US Airline Sentiment.

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2020. Cost-effective Selection of Pretraining Data: A Case Study of Pretraining BERT on Social Media. pages 1675–1681. Association for Computational Linguistics (ACL).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Manas Gaur, Vamsi Aribandi, Amanuel Alambo, Ugur Kursuncu, Krishnaprasad Thirunarayan, Jonathan Beich, Jyotishman Pathak, and Amit Sheth. 2021. Characterization of Time-variant and Time-invariant Assessment of Suicidality on Reddit Using C-SSRS. *PLoS ONE*, 16(5):e0250448.

Debanjan Ghosh, Avijit Vajpayee, and Smaranda Muresan. 2020a. A Report on the 2020 Sarcasm Detection Shared Task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 1–11, Online. Association for Computational Linguistics.

Shalmoli Ghosh, Janardan Misra, Saptarshi Ghosh, and Sanjay Podder. 2020b. Utilizing Social Media for Identifying Drug Addiction and Recovery Intervention. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3413–3422.

Suchin Gururangan, Ana Marasovi´cmarasovi´c, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

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.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics*.

Gondy Leroy, Yang Gu, Sydney Pettygrove, and Margaret Kurzius-Spencer. 2017. Automated Lexicon and Feature Construction Using Word Embedding and Clustering for Classification of ASD Diagnoses Using EHR BT - Natural Language Processing and Information Systems. pages 34–37, Cham. Springer International Publishing.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv*, 1907(11692).

Arjun Magge, Ari Z Klein, Antonio Miranda-Escalada, Mohammed Ali Al-Garadi, Ilseyar Alimova, Zulfat Miftahutdinov, Salvador Lima López, Ivan Flores, Karen O'connor, Davy Weissenbacher, Elena Tutubalina, Juan M Banda, Martin Krallinger, and Graciela Gonzalez-Hernandez. 2021. Overview of the Sixth Social Media Mining for Health Applications (#SMM4H) Shared Tasks at NAACL 2021.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 Task 1: Affect in Tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020a. BERTweet: A Pre-trained Language Model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Dat Quoc Nguyen, Thanh Vu, Afshin Rahimi, Mai Hoang Dao, Linh The Nguyen, and Long Doan. 2020b. WNUT-2020 Task 2: Identification of Informative COVID-19 English Tweets. In *Online*, pages 314–318. Association for Computational Linguistics (ACL).

Dat Quoc Nguyen, Thanh Vu, Afshin Rahimi, Mai Hoang Dao, Linh The Nguyen, and Long Doan. 2020c. WNUT-2020 Task 2: Identification of Informative COVID-19 English Tweets. In *Proceedings of the 6th Workshop on Noisy User-generated Text*.

Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2021. eRisk 2021: Pathological Gambling, Self-harm and Depression Challenges. In *Advances in Information Retrieval*, pages 650–656, Cham. Springer International Publishing.

Romain Paulus and Jeffrey Pennington. Script for Preprocessing Tweets.

Daniel Preoţiuc-Pietro, H Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 2016. Modelling Valence and Arousal in Facebook Posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15.

A. Sarker, M. Belousov, J. Friedrichs, K. Hakala, S. Kiritchenko, F. Mehryary, S. Han, T. Tran, A. Rios, R. Kavuluru, B. De Bruijn, F. Ginter, D. Mahata, S.M. Mohammad, G. Nenadic, and G. Gonzalez-Hernandez. 2018a. Data and Systems for Medication-Related Text Classification and Concept Normalization From Twitter: Insights From the Social Media Mining for Health (SMM4H)-2017 Shared Task. *Journal of the American Medical Informatics Association*, 25(10).

A. Sarker and G. Gonzalez. 2015a. Portable Automatic Text Classification for Adverse Drug Reaction Detection via Multi-Corpus Training. *Journal of Biomedical Informatics*, 53.

Abeed Sarker, Mohammed Ali Al-Garadi, Yuan-Chi Yang, Sahithi Lakamana, Jie Lin, Sabrina Li, Angel Xie, Whitney Hogg-Bremer, Mylin Torres, Imon Banerjee, and Abeed Sarker. 2020. Automatic Breast Cancer Survivor Detection from Social Media for Studying Latent Factors Affecting Treatment Success. *medRxiv*.

Abeed Sarker, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, Berry de Bruijn, Filip Ginter, Debanjan Mahata, Saif M Mohammad, Goran Nenadic, and Graciela Gonzalez-Hernandez. 2018b. Data and Systems for Medication-Related Text Classification and Concept Normalization from Twitter: Insights from the Social Media Mining for Health (SMM4H)-2017 Shared Task. *Journal of the American Medical Informatics Association*, 25(10):1274–1283.

Abeed Sarker and Graciela Gonzalez. 2015b. Portable Automatic Text Classification for Adverse Drug Reaction Detection via Multi-Corpus Training. *J. of Biomedical Informatics*, 53(C):196–207.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

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

Guangyi Xu. 2022. The dilemma and prospects of deep learning. In *2022 IEEE International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA)*, pages 1196–1199.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*.

A Appendix

The full details about the benchmarking tasks and classification results are shown in Table 3.

ID	Task	Source	Evaluation metric	TRN	TST	L	IAA
1	ADR Detection (Sarker and Gonzalez, 2015b)	Twitter	P_F_1	4318	1152	2	0.71
2	Breast Cancer (Sarker et al., 2020)	Twitter	P_F_1	3513	1204	2	0.85
3	NPMU characterization (Ali Al-Garadi et al., 2020)	Twitter	P_F_1*	11829	3271	4	0.86
4	WNUT-20-task2 (COVID-19 tweet detection) (Nguyen et al., 2020c)	Twitter	P_F_1	6238	1000	2	0.8
5	SMM4H-17-task1 (ADR detection) (Sarker et al., 2018b)	Twitter	P_F_1	5340	6265	2	0.69
6	SMM4H-17-task2 (medication consumption) (Sarker et al., 2018b)	Twitter	M_F_1	7291	5929	3	0.88
7	SMM4H-21-task1 (ADR detection) (Magge et al., 2021)	Twitter	P_F_1	15578	913	2	-
8	SMM4H-21-task3a (regimen change on Twitter) (Magge et al., 2021)	Twitter	P_F_1	5295	1572	2	-
9	SMM4H-21-task3b (regimen change on WebMD) (Magge et al., 2021)	WebMD	P_F_1	9344	1297	2	-
10	SMM4H-21-task4 (adverse pregnancy outcomes) (Magge et al., 2021)	Twitter	P_F_1	4926	973	2	0.9
11	SMM4H-21-task5 (COVID-19 potential case) (Magge et al., 2021)	Twitter	P_F_1	5790	716	2	0.77
12	SMM4H-21-task6 (COVID-19 symptom) (Magge et al., 2021)	Twitter	M_F_1	8188	500	3	-
13	SMM4H-22-task9 (self-reporting exact age) (Weissenbacher et al., 2022)	Reddit	M_F_1	7165	1000	2	-
14	Suicidal Ideation Detection (Gaur et al., 2021)	Reddit	M_F_1	1695	553	6	0.88
15	Drug Addiction and Recovery Intervention (Ghosh et al., 2020b)	Reddit	M_F_1	2032	601	5	-
16	eRisk-21-task1 (Signs of Pathological Gambling) (Parapar et al., 2021)	Reddit	P_F_1	1511	481	2	-
17	eRisk-21-task2 (Signs of Self-Harm) (Parapar et al., 2021)	Reddit	P_F_1	926	284	2	-
18	Sentiment Analysis (Food Allergy) (Carrillo-de Albornoz et al., 2018)	MedHelp	M_F_1	618	191	3	0.75
19	Sentiment Analysis (Crohn'S Disease) (Carrillo-de Albornoz et al., 2018)	MedHelp	M_F_1	1056	317	3	0.72
20	Sentiment Analysis (Breast Cancer) (Carrillo-de Albornoz et al., 2018)	MedHelp	M_F_1	551	161	3	0.75
21	Factuality Classification (Food Allergy) (Carrillo-de Albornoz et al., 2018)	MedHelp	M_F_1	580	159	3	0.73
22	Factuality Classification (Crohn'S Disease) (Carrillo-de Albornoz et al., 2018)	MedHelp	M_F_1	1018	323	3	0.75
23	Factuality Classification(Breast Cancer) (Carrillo-de Albornoz et al., 2018)	MedHelp	M_F_1	524	161	3	0.75
24	OLID-1 (Zampieri et al., 2019)	Twitter	M_F_1	11916	860	2	-
25	OLID-2 (Zampieri et al., 2019)	Twitter	M_F_1	11916	240	2	-
26	OLID-3 (Zampieri et al., 2019)	Twitter	M_F_1	11916	213	3	-
27	TRAC-1-1 (Kumar et al., 2018)	Facebook	M_F_1	11999	916	3	-
28	TRAC-1-2 (Kumar et al., 2018)	Twitter	M_F_1	11999	1257	3	-
29	TRAC2-1 (Bhattacharya et al., 2020)	Youtube	M_F_1	4263	1200	3	-
30	TRAC2-2 (Bhattacharya et al., 2020)	Youtube	M_F_1	4263	1200	2	-
31	SemEval-2018 Task 1-4 (Mohammad et al., 2018)	Twitter	PRS	1182	938	8	
32	SemEval-2018 Task 1-2-1 (Mohammad et al., 2018)	Twitter	PRS	1701	1002	4	0.9
33	SemEval-2018 Task 1-2-2 (Mohammad et al., 2018)	Twitter	PRS	1616	1105	4	0.91
34	SemEval-2018 Task 1-2-3 (Mohammad et al., 2018)	Twitter	PRS	1533	975	4	0.83
35	SemEval-2018 Task 1-2-4 (Mohammad et al., 2018)	Twitter	PRS	2252	986	4	0.85
36	Valence CLS (Preoţiuc-Pietro et al., 2016)	Facebook	PRS	2066	604	9	0.77
37	Arousal CLS (Preoţiuc-Pietro et al., 2016)	Facebook	PRS	2088	590	9	0.83
38	Sarcasm-FigLang-Reddit (Ghosh et al., 2020a)	Reddit	M_F_1	3960	1800	2	-
39	Sarcasm-FigLang-Twitter (Ghosh et al., 2020a)	Twitter	M_F_1	4500	1800	2	-
40	Airline (sentiment analysis) (Crowdflower, 2016)	Twitter	M_F_1	10493	2957	3	-

Table 3: Details of the classification tasks and the data statistics. P_F_1 denotes the F_1-score for the positive class, M_F_1 denotes the micro-averaged F_1-score among all the classes, and PRS denotes Pearson correlation coefficient. *For NPMU, P_F_1 denotes the F_1-score of the non-medical use class. TRN, TST, and L denote the training set size, the test set size, and the number of classes, respectively. IAA is the inter-annotator agreement, where Task 4 used Fleiss' K, Task 14 used Krippendorff's alpha, Task 18-23 provided IAA but did not mention the coefficient they used, and other tasks used Cohen's Kappa.