Benchmarking of Transformer-Based Pre-Trained Models on Social Media Text Classification Datasets

Yuting Guo*[1], Xiangjue Dong*[1], Mohammed Ali Al-Garadi[2],
Abeed Sarker[2], Cécile Paris[3], Diego Mollá-Aliod[4]
[1]Department of Computer Science, Emory University, Atlanta, GA, USA
[2]Department of Biomedical Informatics, Emory University, Atlanta, GA, USA
[3]CSIRO Data61, Sydney, Australia
[4]Department of Computing, Macquarie University, Sydney, Australia
{yuting.guo, xiangjue.dong, m.a.al-garadi, abeed.sarker}@emory.edu
cecile.paris@data61.csiro.au, diego.molla-aliod@mq.edu.au

Abstract

Free text data from social media is now widely used in natural language processing research, and one of the most common machine learning tasks performed on this data is classification. Generally speaking, performances of supervised classification algorithms on social media datasets are lower than those on texts from other sources, but recently-proposed transformer-based models have considerably improved upon legacy state-of-the-art systems. Currently, there is no study that compares the performances of different variants of transformer-based models on a wide range of social media text classification datasets. In this paper, we benchmark the performances of transformer-based pre-trained models on 25 social media text classification datasets, 6 of which are health-related. We compare three pre-trained language models, RoBERTa-base, BERTweet and ClinicalBioBERT in terms of classification accuracy. Our experiments show that RoBERTa-base and BERTweet perform comparably on most datasets, and considerably better than ClinicalBioBERT, even on health-related datasets.

1 Introduction

Transformer-based pre-trained language models have proven to be effective for many natural language processing (NLP) tasks, such as text classification and question answering, and they have enabled systems to outperform previous state-of-the-art approaches. A prime example of such language representation models is Bidirectional Encoder Representations from Transformers (BERT), which was pre-trained on the Book Corpus and English Wikipedia (Devlin et al., 2019). Since it was proposed, many efforts have attempted to improve upon it, and common strategies for doing so are to use more data and train longer (Liu et al., 2019), or to pre-train from scratch on domain-specific data (Gu et al., 2020). Multiple variants of transformer-based models have been proposed, but there is currently limited information available about how the variants directly compare on a set of similar tasks.

In this paper, we focus on text from a specific source, namely, social media, and the common task of text classification. We compare the performances of three pre-training methods. We chose text classification as our target task because it is perhaps the most common NLP-related machine learning task, and most of the publicly-available annotated datasets were prepared for it. We included 25 social media classification datasets, 6 of which are health-related. We compared three transformer-based models—RoBERTa-base (Liu et al., 2019), BERTweet (Nguyen et al., 2020a), and ClinicalBioBERT (Alsentzer et al., 2019). Our experiments show that RoBERTa-base and BERTweet perform comparably and are considerably better than ClinicalBioBERT. In addition to comparing the performances of the models on all the datasets, we analyzed the differences in performances between domain-specific (medical), source-specific (social media), and generic pre-trained models. Our empirical analyses suggest that RoBERTa-base can capture general text characteristics, while BERTweet can capture source-specific knowledge, and pre-training on large-scale source-specific data can improve the capabilities of models to capture general text features, potentially benefiting downstream source-specific tasks.

2 Related Work

The most relevant and recent related works are those by Peng et al. (2019) and Gu et al. (2020). Peng et al. (2019) proposed the Biomedical Language Understanding Evaluation (BLUE) benchmark for the biomedical domain. The evaluations include five tasks with ten datasets covering both biomedical and clinical texts. The specific tasks include named entity recognition, text classification and relation extraction. Gu et al. (2020) proposed the Biomedical Language Understanding and Reasoning Benchmark (BLURB) for PubMed-based biomedical NLP applications, with 13 biomedical NLP datasets in six tasks. To the best of our knowledge, there is no existing work that attempts to perform similar benchmarking for transformer-based approaches on social media data, and the results reported in this paper follow on the footsteps of the benchmarks referenced above.

Recent attempts at adaptation of transformer-based models are also relevant to our current work, since we wanted to include a domain-adapted and a source-adapted model in our comparisons. Many domain adaptation efforts have been reported in the literature. BioBERT—generated by pre-training BERT on biomedical corpora (*e.g.*, PubMed abstracts)—was demonstrated to outperform BERT on three representative biomedical text mining tasks (Lee et al., 2019). Alsentzer et al. (2019) attempted to further adapt pre-trained models for clinical text by training BioBERT on clinical notes, resulting in the ClinicalBioBERT model. We included ClinicalBioBERT as an example of a domain-adapted pre-trained model in our comparisons. For source-adaptation (social media text), Nguyen et al. (2020a) proposed BERTweet by pre-training BERT on a large set of English tweets. We include BERTweet in our comparisons as an example of a source-adapted model.

3 Methods

3.1 Model Architecture

We focus solely on benchmarking systems for social media text classification datasets in this paper. The overall framework of our classification model is shown in Figure 1. It consists of an encoder, a pooling layer, a linear layer, and an output layer with Softmax activation. The encoder converts each token in a document into a embedding matrix, and the pooling layer generates a document embedding e_d by averaging the word embeddings.[1] The document embedding is then fed into the linear layer and the output layer. The output is a probability value between 0 and 1, which is used to compute a logistic loss during the training phase, and the class with the highest probability is chosen in the inference phase. We use the encoders from recent pre-trained deep language models that are trained on different corpora and pre-training tasks to convert documents into embeddings, as described in Section 3.2.

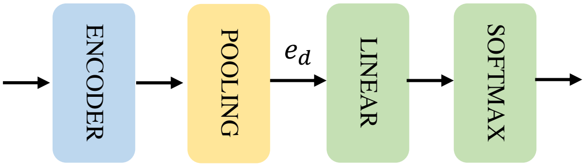

Figure 1: The overall framework of our model.

3.2 Document Encoder

RoBERTa: A BERT variant named RoBERTa was released by Liu et al. (2019) with the same model architecture of BERT but with improved performance, achieved by training the model longer with a larger batch size, on more data, removing the next sentence prediction objective during the pre-training procedure, and applying a dynamic masking technique. We chose RoBERTa-base as the generic or domain-independent encoder in this paper since it outperforms BERT-base and matches the state-of-the-art results of another BERT variant XLNet (Yang et al., 2019) on some NLP tasks.

BERTweet: Nguyen et al. (2020a) developed BERTweet, a pre-trained deep language model with the same model architecture as BERT-base, but using the RoBERTa pre-training procedure on a large scale set of English tweets. Because tweets generally use informal grammar and irregular vocabulary, which are different from traditional text data such as news articles and Wikipedia, BERTweet was an attempt at source adaptation of pre-trained models. BERTweet has been shown to obtain better results than RoBERTa-base on three Tweet NLP tasks—POS tagging, named entity recognition and text classification, illustrating its higher capability of capturing language features of English Tweets compared to RoBERTa-base (Nguyen et al., 2020a).

ClinicalBioBERT: ClinicalBioBERT (Alsentzer et al., 2019), is built by further training of BioBERT (Lee et al., 2019) on clinical notes, and it has been shown to significantly outperform BERT-base on three clinical NLP tasks. This model can generate contextual word embeddings, which are expected to capture clinical knowledge and can benefit the clinical NLP tasks such as natural language inference and entity recognition in the medical domain.

[1] We also experimented with [CLS] embeddings, but did not observe significant performance differences (Appendix A.2).

3.3 Data

We included 25 datasets in our experiments, comprising 6 datasets that were created for health-related tasks such as prescription medication abuse and adverse drug reaction detection, and 19 that were created for non-health-related tasks such as sentiment analysis and offensive language detection. The detailed data descriptions are listed in the Appendix A.1, and the statistics of all datasets are described in Table 1. For data preprocessing, we followed the procedure implemented by the open source tool *preprocess-twitter*,[2] which includes the steps of lowercasing, and normalizing numbers, hashtags, links, capital words and repeated letters.

	Dataset	TRN	TST	L	S
Health	ADR Detection	4318	1152	2	T
	BreastCancer	3513	1204	2	T
	PM Abuse	11829	3271	4	T
	SMM4H-17-task1	5340	6265	2	T
	SMM4H-17-task2	7291	5929	3	T
	WNUT-20-task2	6238	1000	2	T
Non-Health	OLID-1	11916	860	2	T
	OLID-2	11916	240	2	T
	OLID-3	11916	213	3	T
	TRAC-1-1	11999	916	3	F
	TRAC-1-2	11999	1257	3	T
	TRAC-2-1	4263	1200	3	Y
	TRAC-2-2	4263	1200	2	Y
	Sarcasm-1	3960	1800	2	R
	Sarcasm-2	4500	1800	2	T
	CrowdFlower	28707	8101	13	T
	FB-arousal-1	2085	580	9	F
	FB-arousal-2	2088	590	9	F
	FB-valence-1	2064	595	8	F
	FB-valence-2	2066	604	9	F
	SemEval-18-A	1701	1002	4	T
	SemEval-18-F	2252	986	4	T
	SemEval-18-J	1616	1105	4	T
	SemEval-18-S	1533	975	4	T
	SemEval-18-V	1182	938	8	T

Table 1: The statistics of the training (`TRN`) and test (`TST`) set. `L`: #classes; `S`: data sources; `T`: Twitter; `R`: Reddit; `F`: Facebook; `Y`: YouTube.

3.4 Experimental Setup

Following a modified setting from Liu et al. (2019), we performed a limited parameter search with learning rate $\in \{2e-5, 3e-5\}$. We fine-tuned each model for 10 epochs and selected the model that achieves the best metric on the validation set. Each experiment was run three times with different initializations, and the median results of the validation and test sets for each dataset are reported. The rest of hyper-parameters were empirically chosen and are shown in Table 2.

Hyper-parameter		**Hyper-parameter**	
Max sequence size	128	Warmup ratio	0
Batch size	32	Adam epsilon	1e-8

Table 2: Hyper-parameter configurations of all models.

4 Results and Discussion

Table 3 lists the accuracies of all the models on the test sets of the included datasets. In order to compare the statistical significance of differences between the accuracies, we used the McNemar's test to compare the top-2 best models for each dataset. The difference between two models is regarded as statistically significant if the p-value <0.05.

Dataset	RB	BT	CL	p-value
ADR Detection	91.4	**92.7**	90.4	0.11
BreastCancer	**93.9**	93.6	91.2	0.90
PM Abuse	81.4	**82.4**	77.4	0.09
SMM4H-17-task1	**93.6**	93.5	92.7	0.76
SMM4H-17-task2	78.4	**79.7**	75.0	**0.01**
WNUT-20-task2	**89.1**	88.3	86.5	0.48
OLID-1	85.1	**85.2**	83.5	0.90
OLID-2	89.4	**90.0**	89.0	0.73
OLID-3	69.5	**70.0**	66.4	0.73
TRAC-1-1	58.6	**59.2**	55.4	0.76
TRAC-1-2	58.8	**65.8**	58.0	**0.00**
TRAC-2-1	72.8	**73.3**	63.9	1.00
TRAC-2-2	85.8	85.5	**87.2**	0.10
sarcasm-1	67.3	**69.5**	64.6	0.06
sarcasm-2	73.2	**76.1**	68.2	**0.02**
CrowdFlower	39.9	**41.3**	38.8	**0.00**
fb-arousal-1	46.6	45.3	**46.8**	1.00
fb-arousal-2	**54.9**	54.8	54.1	0.92
fb-valence-1	60.2	**64.4**	54.5	0.06
fb-valence-2	**52.8**	52.6	45.9	1.00
SemEval-18-A	52.3	**54.6**	46.0	0.16
SemEval-18-F	**69.3**	67.4	65.3	0.09
SemEval-18-J	47.7	**51.5**	45.3	**0.01**
SemEval-18-S	**54.9**	53.9	48.4	0.42
SemEval-18-V	45.5	**46.6**	36.2	0.56

Table 3: The accuracies of the three transformer-based models on the test splits of our included datasets. `RB`: RoBERTa; `BT`: BERTweet; `CL`: ClinicalBioBERT; `p-value`: McNemar's test p-value. The best result of each dataset and the p-values <0.05 are in boldface.

BERTweet achieves the highest accuracies on 16 out of 25 datasets, including health and non-health-related datasets from Twitter, Facebook, Reddit, and YouTube. The fact that BERTweet performs well on non-tweet datasets suggests that BERTweet can learn some universal characteristics of social media languages by pre-training on tweets.[3] On 5 datasets (specifically, `SMM4H-17-task2`,

[2] `https://nlp.stanford.edu/projects/glove/preprocess-twitter.rb`

[3] Dai et al. (2020) reported a similar finding: a model pre-trained on business reviews (Forum BERT) outperformed one

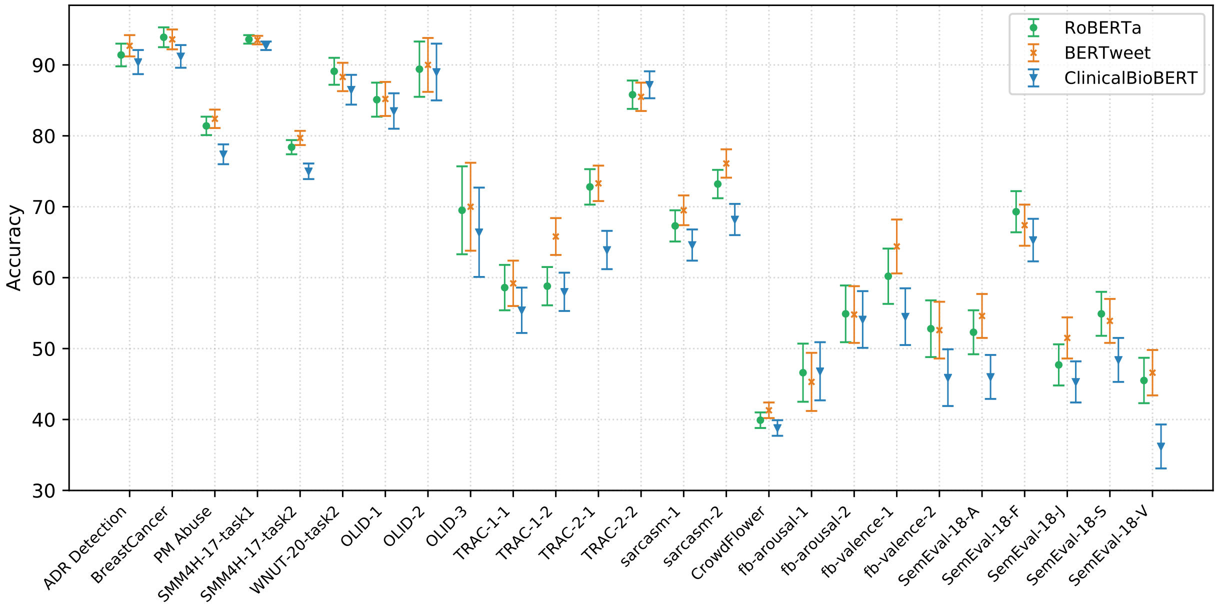

Figure 2: The 95% confidence intervals of three models on our included datasets.

`TRAC-1-2`, `sarcasm-2`, `CrowdFlower`, and `SemEval-18-J`), the best-performing system obtained significantly better results than the next best system, and in all these cases, BERTweet was the winner. There are, however, no significant differences between RoBERTa-base and BERTweet on most datasets, which shows that RoBERTa-base can capture general text features and work well on social media tasks. The differences in the pre-training dataset sizes for RoBERTa-base (160 GB) and BERTweet (80 GB) suggest that pre-training on relatively small source-specific data may effectively benefit the downstream source-specific tasks.

Figure 2 visually illustrates the distribution of the accuracy scores and their 95% confidence intervals for all three models on our included datasets. From the figure, the relative underperformance of the ClinicalBioBERT is evident. ClinicalBioBERT does not appear to capture social media-specific characteristics of the data even for health-related classification datasets, although it is trained on clinical notes. This finding suggests that for social media-specific health-related research tasks, it might be better to choose a source-specific pre-trained model (*e.g.*, BERTweet for social media) rather than a domain-specific one. It is possible that the gap between the language of clinical notes and social media text is large enough to negatively impact the social media text representation capability of the encoder. Moreover, ClinicalBioBERT is trained by continuing the training of BioBERT on a small size of clinical data (about 2 million records), which may have led to the insufficient learning of clinical knowledge. The under-performance of ClinicalBioBERT does not necessarily mean that domain-specialized transformer models are inferior. Our experimental results also suggest that large pre-training data can boost the generalizability of models, while pre-training on small in-domain data may not benefit target tasks within the domain. Based on our findings, for social media text classification datasets, we recommend the use of RoBERTa-base, BERTweet or models pre-trained in similar fashion, and we do not recommend the use of ClinicalBioBERT, even for health-related social media tasks. A major limitation of our current work is that we only evaluated three pre-trained models, and, in the future, we will incorporate other similar models such as Twitter BERT (Dai et al., 2020) and BioBERT (Lee et al., 2019). We will also evaluate models using more metrics, as accuracy can be particularly misleading for imbalanced datasets.

5 Conclusion

We benchmarked the performances of three transformer-based pre-trained models on 25 social media text classification datasets. We found that RoBERTa-base and BERTweet perform similarly on most datasets, consistently outperforming ClinicalBioBERT, even for health-related tasks. Our experiments suggest that for social media-based classification tasks, it might be best to use

pre-trained on tweets (Twitter BERT) on 3 tweet classification tasks.

pre-trained models generated from large social media text. It might be possible to further improve the performance of BERTweet by incorporating data from multiple social networks.

References

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 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).

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. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1675–1681, Online. Association for Computational Linguistics.

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*.

Debanjan Ghosh, Avijit Vajpayee, and Smaranda Muresan. 2020. 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.

Yu Gu, Robert Tinn, Hao Cheng, M. Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ArXiv*, abs/2007.15779.

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*, 36(4):1234–1240.

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).

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 *Proceedings of the 6th Workshop on Noisy User-generated Text*.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

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.

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. 2018.

Dataset	Description	Source
ADR Detection	Detect adverse reaction (ADR) mentioned from text	Sarker and Gonzalez (2015)
BreastCancer	Detect breast cancer patients based on their self-reports	Sarker et al. (2020)
PM Abuse	Identify prescription medication (PM) abuse on tweets	Ali Al-Garadi et al. (2020)[4]
SMM4H-17-task1	Detect adverse reaction (ADR) mentioned from text	Sarker et al. (2018)
SMM4H-17-task2	Identify medication consumption from medication-mentioning tweets	
WNUT-20-task2	Identify informative COVID-19 related tweets	Nguyen et al. (2020b)
OLID-1	Identify offensive language from tweets	Zampieri et al. (2019)
OLID-2		
OLID-3		
TRAC-1-1	Detect aggressive information in social media	Kumar et al. (2018)
TRAC-1-2		
TRAC-2-1	Detect aggressive language on social media text	Bhattacharya et al. (2020)
TRAC-2-2		
sarcasm-1	Binary emotion classification of sarcasm	Ghosh et al. (2020)
sarcasm-2		
CrowdFlower	Multiclass emotion classification	Web[5]
fb-arousal-1	Classify the level of arousal	Preoţiuc-Pietro et al. (2016)
fb-arousal-2		
fb-valence-1	Classify the level of valence	
fb-valence-2		
SemEval-18-A	Emotion intensity ordinal classification of anger	Mohammad et al. (2018)
SemEval-18-F	Emotion intensity ordinal classification of fear	
SemEval-18-J	Emotion intensity ordinal classification of joy	
SemEval-18-S	Emotion intensity ordinal classification sadness	
SemEval-18-V	Valence ordinal classification	

Table A.1: Data descriptions.

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

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc.

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

A.1 Data Descriptions

Table A.1 provides a short description about the classification task focuses. The datasets that do not provide a split of train/dev/test sets are split into a training set and a test set using a 80/20 rate. For WNUT-20-task2, the results on the validation set was reported because the test set was not released.

A.2 Pooling Strategy Comparison

Table A.2 shows the results of taking `[CLS]` embeddings as document embeddings.

Dataset	RB		BT		CL	
	C	M	C	M	C	M
ADR Detection	91.7	91.4	90.4	**92.7**	90.8	90.4
BreastCancer	**94.1**	93.9	93.4	93.6	90.8	91.2
PM Abuse	81.1	81.4	81.9	**82.4**	77.4	77.4
SMM4H-17-task1	**93.6**	**93.6**	93.2	93.5	92.3	92.7
SMM4H-17-task2	78.9	78.4	79.1	**79.7**	74.3	75.0
WNUT-20-task2	**89.7**	89.1	88.3	88.3	85.8	86.5
OLID-1	**85.5**	85.1	84.7	85.2	83.4	83.5
OLID-2	89.2	89.4	**90.6**	90.0	89.2	89.0
OLID-3	68.5	69.5	**71.4**	70.0	67.8	66.4
TRAC-1-1	57.5	58.6	**59.2**	**59.2**	52.2	55.4
TRAC-1-2	58.6	58.8	**65.8**	**65.8**	57.4	58.0
TRAC2-1	**75.1**	72.8	63.3	73.3	66.3	63.9
TRAC2-2	85.4	85.8	83.9	85.5	**87.6**	87.3
CrowdFlower	39.8	39.9	35.0	**41.3**	38.8	38.8
fb-arousal-1	45.8	46.6	45.6	45.3	45.7	**46.8**
fb-arousal-2	54.6	**54.9**	52.9	54.8	52.4	54.1
fb-valence-1	59.5	60.2	60.5	**64.4**	52.9	54.5
fb-valence-2	**53.6**	52.8	52.6	52.6	44.9	45.9
sarcasm-1	66.3	67.3	**71.4**	69.5	64.8	64.6
sarcasm-2	73.2	73.3	**76.2**	76.1	68.0	68.2
SemEval-18-task-A	55.4	52.3	**60.8**	54.6	48.9	46.0
SemEval-18-task-F	49.4	47.7	43.4	**51.5**	45.1	45.3
SemEval-18-task-J	53.7	**54.9**	53.9	53.9	49.7	48.4
SemEval-18-task-S	68.2	**69.3**	64.2	67.4	65.9	65.3
SemEval-18-task-V	45.7	45.5	38.3	**46.6**	36.4	36.2

Table A.2: The accuracies of taking different pooling strategies on the test sets. `C`: [CLS] emebddings; `M`: mean word embeddings. The best results on each dataset are in boldface.

[5] https://projectreporter.nih.gov/project_info_description.cfm?aid=9577760

[5] https://data.world/crowdflower/sentiment-analysis-in-text