zydhjh4593@SMM4H'22: A Generic Pre-trained BERT-based Framework for Social Media Health Text Classification

Chenghao Huang[1], Xiaolu Chen[1], Yuxi Chen[1], Yutong Wu[2],
Weimin Yuan[1], Yan Wang[1], and Yanru Zhang[1,3]

[1]University of Electronic Science and Technology of China
[2]University of Nottingham Ningbo China
[3]Shenzhen Institute of Advanced Study, UESTC
Email: zydhjh4593@gmail.com yanruzhang@uestc.edu.cn

Abstract

This paper describes our proposed framework for the 10 text classification tasks of Task 1a, 2a, 2b, 3a, 4, 5, 6, 7, 8, and 9, in the Social Media Mining for Health (SMM4H) 2022. According to the pre-trained BERT-based models, various techniques, including regularized dropout, focal loss, exponential moving average, 5-fold cross-validation, ensemble prediction, and pseudo-labeling, are applied for further formulating and improving the generalization performance of our framework. In the evaluation, the proposed framework achieves the **1st place** in Task 3a with a 7% higher F_1-score than the median, and obtains a 4% higher averaged F_1-score than the median in all participating tasks except Task 1a.

1 Introduction

Social media platforms such as Twitter and Reddit are full of various statements made by users, and there is abundant social media data that can be used for mining health-related information. Natural language processing (NLP) plays an important role in social media data manipulation, which helps solve issues including informal expressions, misspelling of terms, noise, and multiple languages in the tweets. The intersection of social media data mining and health applications is concerned in the 7th Social Media Mining for Health Applications (SMM4H) workshop (Weissenbacher et al., 2022). Our team participates in 10 classification tasks of the 7th SMM4H shared tasks, which are Task 1a (Magge et al., 2021), 2a, 2b (Davydova and Tutubalina, 2022), 3a, 4-9.

In this paper, a generic framework is formulated to fine-tune pre-trained bidirectional encoder representations from transformers (BERT)-based models on text classification datasets of SMM4H 2022 shared tasks. For obtaining higher accuracy, we adopt 5-fold cross-validation (CV) to get 5 models and use an ensemble learning technique to obtain the result. Furthermore, considering the evaluation performance on test datasets, we fine-tune our models using regularized dropout (R-drop) (Wu et al., 2021) and exponential moving average (EMA) to increase the models' generalization capability. To deal with data imbalance, focal loss (Lin et al., 2017) is adopted to assign higher weights to samples with minority classes. Additionally, pseudo-labeling (Lee et al., 2013) is adopted as an attempt for improvement. As a result of applying the above techniques, our framework obtains a 4% higher averaged F_1-score than the median in all participating tasks except Task 1a, and a 7% higher F_1-score than the median in Task 3a.

2 Preliminaries

2.1 Problem Description and Datasets

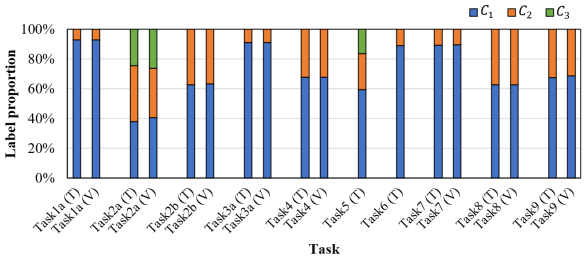

Figure 1: The proportion of different labels in the SMM4H 2022 Task 1a, 2a, 2b, 3a, 4-9, where T and V represent each corresponding task's training dataset and validation dataset, respectively. For Task 1a, C_1 is *noADE* and C_2 is *ADE*. For Task 2a, C_1, C_2, and C_3 are *FAVOR*, *NONE*, and *AGAINST*, respectively. For Task 5, C_1, C_2, and C_3 are *lit-news_mention*, *non-personal_report*, and *self_report*. For Task 6, C_1 is *Vaccine_chatter* and C_2 is *Self_reports*. C_1 and C_2 of the rest tasks are 0 and 1, respectively.

We participate in 10 text classification tasks, including 8 binary classification tasks and 2 three-way classification tasks summarized as follows.

Binary classification tasks:

- **Task 1a** aims to classify if there is an adverse

Proceedings of the 29th International Conference on Computational Linguistic, pages 11–15
October 12–17, 2022.

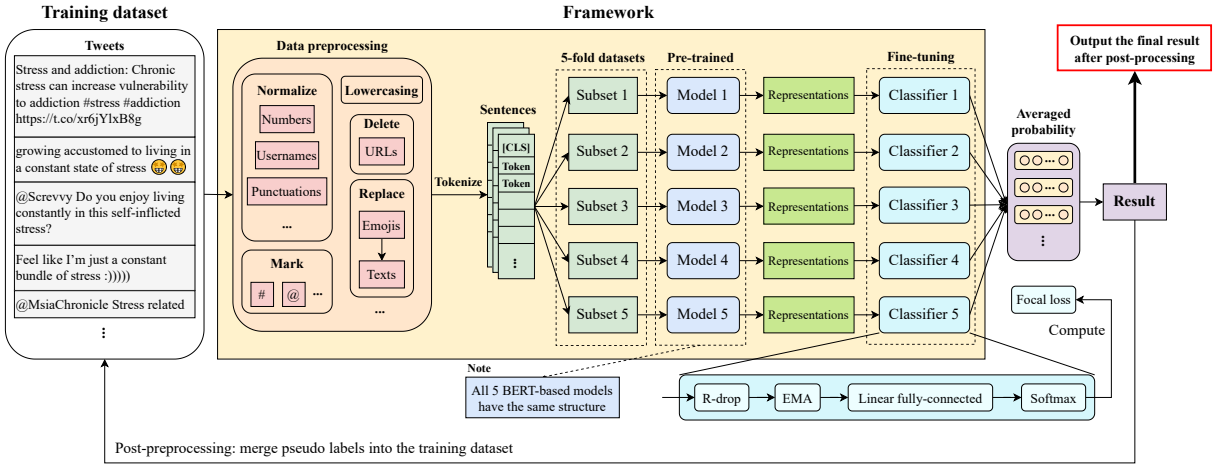

Figure 2: Our framework architecture for the SMM4H 2022 Task 1a, 2a, 2b, 3a, 4-9.

drug event (ADE) in an English tweet.

- **Task 2b** aims to classify if at least one premise/argument is mentioned in a COVID-related English tweet.
- **Task 3a** aims to classify if there is a medication treatment change in an English tweet.
- **Task 4** aims to determine if a self-reporting age is exact in an English tweet.
- **Task 6** aims to classify if there is a COVID-19 vaccination confirmation or just related chatter in an English tweet.
- **Task 7** aims to classify if intimate partner violence exists in an English tweet.
- **Task 8** aims to classify if chronic stress exists in an English tweet (Yang et al., 2022).
- **Task 9** aims to determine if a self-reporting age is exact in an English posting on Reddit.

Three-way classification tasks:

- **Task 2a** aims to determine if a stance is positive, negative, or neutral in an English tweet.
- **Task 5** aims to distinguish if a Spanish tweet about COVID-19 symptoms is a self-report, a non-personal report, or a literature/news mention.

For further analysis, we count the label proportions of each task, illustrated in Figure 1, where C_1, C_2, and C_3 represents different labels in each task. Note that validation datasets of Task 5 and 6 have no labels before the release of the final results.

2.2 Related Work

In recent years, BERT (Devlin et al., 2018), a pre-trained transformer-based model for language representation, has become increasingly successful in text classification. An upgraded version of BERT called RoBERTa (Liu et al., 2019) produces better processing outcomes for NLP tasks. The first pre-trained language model for English tweets, BERTweet (Nguyen et al., 2020), which is built on BERT and was trained using the RoBERTa pre-training procedure, performs better on Twitter NLP tasks. In previous SMM4H workshops, a number of participants used BERT-based models with encouraging results in text classification tasks (Valdes et al., 2021; Ramesh et al., 2021; Sakhovskiy et al., 2021).

3 Framework

A generic framework is formulated for all classification tasks, which is shown in Figure 2. Before model training, simple data preprocessing is conducted. Then, to encode tweet texts, BERT-based classifiers are trained using 5-fold CV. During the inference phase, ensemble learning is applied for more accurate prediction over individual classifiers. In an attempt to further improve the performance, pseudo-labeling based on the inferred test dataset is used.

3.1 Preprocessing

Firstly, to reduce tweets' noise, data preprocessing methods, such as lowercasing, deleting URLs, replacing emojis with their text strings, normalization of numbers, punctuations, and usernames, and marking special characters (#, @, ...), are generally applied on tweets in all participating tasks. Then the training dataset is shuffled and divided evenly into 5 subsets for 5-fold CV during the training

phase. After training, 5 models are obtained for the ensemble prediction during the inference phase. Particularly, tweets in validation datasets of Task 5 and 6 have no labels. For each of them, we simply use the training dataset for 5-fold CV.

3.2 Model Training and Inference

Considering the increasing prevalence of BERT-based model usage in text classification, we propose to adopt pre-trained BERT-based models and fine-tune them using different downstream tasks, i.e., shared tasks in SMM4H 2022. By inputting tokenized tweet texts and other information, such as claims in Task 2a and 2b, to BERT-based models, token representations are obtained. Then, we pass these representations through a linear fully-connected layer and a softmax activation function to get the probability of each class. For acquiring more accurate and robust predictions, 5 models are individually trained through 5-fold CV.

During the training phase, R-drop is adopted to alleviate the model overfitting. By constraining the parameter space through symmetric Kullback-Leibler divergence, R-drop is also capable of reducing randomness caused by the traditional dropout technique. Furthermore, in order to further raise the performance on the test dataset, EMA, which averages model parameters in the latest few steps, is applied to improve the generalization capability of our models. In addition, Figure 1 illustrates that the label proportions show different degrees of imbalance. As BERT-based models and fully-connected layers are constructed for our framework, focal loss, which is designed for neural networks, is adopted to prioritize samples with minority labels that are hard to be classified.

During the inference phase, an ensemble technique is used to combine 5 models' results for better performance. The outputs of the 5 models, i.e., all 5 probability vectors, are averaged. Then, the result is obtained according to the highest values in the averaged probability vector.

3.3 Post-processing

To further improve our models' performance, pseudo-labeling is adopted in the post-processing phase. After inference, prediction on test dataset is set as the pseudo labels. Then, training dataset, validation dataset, and test dataset are merged into an entire dataset for another training. Finally, the latest 5 models conduct ensemble inference and output the final prediction for submission.

4 Experiments

4.1 Setup

To compare the performance of different pre-trained models, 3 BERT-based models pre-trained on English corpora are experimented with: (i) BERT, (ii) RoBERTa, and (iii) BERTweet. Note that validation datasets' labels of Task 5 and 6 are released as well as the results. For the offline evaluation, we conduct inference on the validation datasets to sum the micro F_1-score and the macro F_1-score of each task and calculate the mean value. All simulations and approaches are coded in Python and conducted on a personal computer with an NVIDIA GeForce RTX 3090 Ti GPU. Adam optimizer is used (Kingma and Ba, 2014). The hyperparameters are listed in Appendix A.

4.2 Online Evaluation

As shown in Table 1, our framework obtains a 4% higher averaged F_1-score than the median in all participating tasks except Task 1a, and a 7% higher F_1-score than the median in Task 3a. Besides, our framework's precision values of Task 3a and 4 are 6% and 5% higher than the median, respectively. Our framework's recall values of Task 3a, 7, and 8 are 7%, 8%, and 10% higher than the median, respectively. In Task 1a, though our framework's F_1-score and recall value are 2.5% and 7% lower than the mean, respectively, our precision is about 8% higher than the mean.

Task	F_1-score		Precision		Recall	
	Ours	Median	Ours	Median	Ours	Median
2a	**0.586**	0.550	/	/	/	/
2b	**0.701**	0.647	/	/	/	/
3a	**0.655**	0.586	**0.682**	0.617	**0.630**	0.557
4	**0.914**	0.869	**0.921**	0.869	**0.908**	0.889
5	**0.860**	0.840	**0.860**	0.840	**0.860**	0.840
6	**0.800**	0.770	**0.900**	0.900	**0.710**	0.680
7	**0.795**	0.763	**0.795**	0.790	**0.795**	0.716
8	**0.792**	0.750	**0.734**	0.720	**0.859**	0.760
9	**0.891**	0.891	0.893	**0.896**	0.889	**0.919**

Table 1: Our results and the median results on the test datasets of all participating tasks, where the best results are in bold.

4.3 Offline Evaluation

We evaluate the effects of different pre-trained models and techniques on our framework. As shown in Table 2, when not using pseudo labels, our framework achieves the highest F_1-score in Task 1a, 2a, 2b, and 3a. In the rest of the tasks, the use of pseudo labels shows the best performance.

Models	Task1a	Task2a	Task2b	Task3a	Task4	Task5	Task6	Task7	Task8	Task9
BERT	0.558	0.534	0.660	0.587	0.824	0.806	0.737	0.730	0.732	0.814
RoBERTa	0.585	0.551	0.682	0.619	0.858	0.833	0.764	0.755	0.759	0.847
BERTweet (BT)	0.606	0.574	0.707	0.645	0.897	0.869	0.798	0.784	0.791	0.883
BT+OS w/o FL	0.579	0.566	0.699	0.611	0.882	0.854	0.757	0.749	0.776	0.874
BT+Dropout	0.639	0.618	0.751	0.692	0.945	0.908	0.852	0.822	0.827	0.930
BT+RD	0.654	0.630	0.761	0.705	0.951	0.917	0.862	0.833	0.836	0.928
BT+RD+EMA	**0.662**	**0.642**	**0.768**	**0.718**	0.955	0.924	0.872	0.841	0.843	0.934
BT+RD+EMA+PS	0.638	0.625	0.752	0.713	**0.971**	**0.932**	**0.876**	**0.842**	**0.851**	**0.945**

Table 2: Offline F_1-scores of different models with various techniques on the validation dataset of each participating task, where oversampling, focal loss, R-drop, exponential moving average, and pseudo-labeling are abbreviated as OS, FL, RD, EMA, and PS, respectively, and the best results are in bold. Note that 5-fold CV and focal loss are applied in all models.

5 Result Analysis

Through the results shown in Table 2, we can see that BERTweet outperforms BERT and RoBERTa in all participating tasks, indicating the benefit of pre-training on the right corpora. Because the data of Task 9 is from Reddit, our framework's online F_1-score, precision, and recall in Task 9 are in line with the median, shown in Table 1. Besides, the performance on Task 5 is slightly worse than other tasks, which can be attributed to the Spanish data.

In terms of training techniques, the 3rd and 5th rows of Table 2 show that in all participating tasks, the application of dropout improves around 5% on the BERTweet model, indicating that dropout is crucial to help BERT-based models avoid overfitting. Moreover, the effectiveness of R-drop is proved as the performance of R-drop is around 1% higher than the traditional dropout. In addition, EMA helps with the fine-tuning convergence, improving around 0.7% on R-drop-based BERTweet in all participating tasks.

To verify the effectiveness of focal loss, we compare focal loss with oversampling, which randomly duplicates samples with minority classes until achieving data balance. As shown in the 3rd and 4th rows of Table 2, F_1-scores of using focal loss are all higher than oversampling, especially on validation datasets of Task 1a, 3a, 6, and 7, where labels are significantly imbalanced. The reason is that though oversampling enlarges the sizes of training data, overfitting may occur in samples with minority classes. Besides, as the dataset sizes in all participating tasks are not large, undersampling, which will lead to overfitting by reducing the training datasets, is discarded.

Additionally, Table 2 shows that pseudo-labeling is not effective on all participating tasks. We count labeled and unlabeled data, which are shown in Figure 3. As for Task 1a, 2a, 2b, and 3a, where the prediction performance is not as expected, and the size of each test dataset is fairly large. In such a situation, pseudo-labeling will weaken the performance of models. On the contrary, in the other tasks, pseudo-labeling is effective. Therefore, it is concluded that pseudo-labeling only contributes to models which have already obtained outstanding performance.

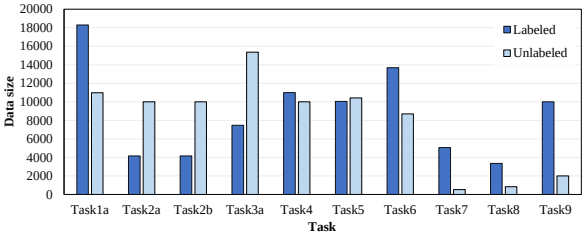

Figure 3: The statistics of labeled data and unlabeled data in the SMM4H 2022 Task 1a, 2a, 2b, 3a, 4-9.

6 Conclusion

In this work, A generic framework based on pre-trained BERT-based models is formulated for 10 text classification tasks, which are Task 1a, 2a, 2b, 3a, 4-9 in the SMM4H 2022. Through 5-fold CV, 5 models are trained and fine-tuned with R-drop, EMA, and focal loss. Then, during the inference, ensemble prediction from 5 models is outputted. In addition, pseudo-labeling is applied for the better fitting capability of our models. As a result, our framework obtains a 4% higher averaged F_1-score than the median in all participating tasks except Task 1a, and achieves the 1st place in Task 3a with a 7% higher F_1-score than the median.

For future work, in order to sufficiently utilize labeled data, contrastive learning (Khosla et al., 2020) will be studied between tweet texts and other information, such as claims in Task 2.

References

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

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

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

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*.

Arjun Magge, Elena Tutubalina, Zulfat Miftahutdinov, Ilseyar Alimova, Anne Dirkson, Suzan Verberne, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021. Deepademiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on twitter. *Journal of the American Medical Informatics Association*, 28(10):2184–2192.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*.

Sidharth Ramesh, Abhiraj Tiwari, Parthivi Choubey, Saisha Kashyap, Sahil Khose, Kumud Lakara, Nishesh Singh, and Ujjwal Verma. 2021. BERT based transformers lead the way in extraction of health information from social media. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 33–38, Mexico City, Mexico. Association for Computational Linguistics.

Andrey Sakhovskiy, Zulfat Miftahutdinov, and Elena Tutubalina. 2021. KFU NLP team at SMM4H 2021 tasks: Cross-lingual and cross-modal BERT-based models for adverse drug effects. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 39–43, Mexico City, Mexico. Association for Computational Linguistics.

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

Davy Weissenbacher, Ari Z. Klein, Luis Gascó, Darryl Estrada-Zavala, Martin Krallinger, Yuting Guo, Yao Ge, Abeed Sarker, Ana Lucia Schmidt, Raul Rodriguez-Esteban, Mathias Leddin, Arjun Magge, Juan M. Banda, Vera Davydova, 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 Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages –.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905.

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

A Hyperparameters

Module	Hyperparameter	Value
Classifier	Batch size	16
	Updata frequency	10 (epochs)
	Learning rate (LR)	1e-4
	LR decay rate	0.99
	Focal loss factor	0.3
	Adam epsilon	1e-6
	EMA start	0
	EMA decay	0.99
	R-drop rate	0.5
	R-drop weight	5
BERT	Embedding size	768
	Hidden size	768
	Sequence length	128
	Learning rate	2e-5

Table 3: Hyperparameters of the proposed framework.