Pre-trained Transformer-based Classification and Span Detection Models for Social Media Health Applications

Yuting Guo and **Yao Ge**
Computer Science
Emory University
Atlanta GA 30322, USA
yuting.guo@emory.edu
yao.ge@emory.edu

Mohammed Al-Garadi and **Abeed Sarker**
Biomedical Informatics
Emory University
Atlanta GA 30322, USA
maalgar@emory.edu
abeed@dbmi.emory.edu

Abstract

This paper describes our approach for six classification tasks (Tasks 1a, 3a, 3b, 4 and 5) and one span detection task (Task 1b) from the Social Media Mining for Health (SMM4H) 2021 shared tasks. We developed two separate systems for classification and span detection, both based on pre-trained Transformer-based models. In addition, we applied oversampling and classifier ensembling in the classification tasks. The results of our submissions are over the median scores in all tasks except for Task 1a. Furthermore, our model achieved first place in Task 4 and obtained a 7% higher F_1-score than the median in Task 1b.

1 Introduction

Social media platforms such as Twitter have been widely used to share experiences and health information such as adverse drug effects (ADEs), thus attracting an increasing number of researchers to conduct health-related research using this data. However, because social media data consists of user-generated content that is noisy and written in informal language, health language processing with social media data is still challenging. To promote the use of social media for health information extraction and analysis, the Health Language Processing Lab of the University of Pennsylvania organized Social Media Mining for Health Applications (SMM4H) shared tasks. This year, the SMM4H shared tasks included 8 subtasks (Magge et al., 2021). Our team, the Sarker Lab at Emory University, participated in six classification tasks (*i.e.*, Task 1a, 3a, 3b, 4, and 5) and one span detection task (*i.e.*, Task 1b) of the SMM4H 2021 shared tasks. In recent years, Transformer-based models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), whose advantage is modeling of long-range context semantics, revolutionised the field of NLP and achieved state-of-the-art results in different NLP tasks. Encouraged by those successes, we developed separate systems for classification and span detection both based on pre-trained Transformer-based models. We experimented with different Transformer-based model variants, and the model that achieved the best result on the validation set was selected as the final system. In addition, we performed undersampling and oversampling to address the problem of data imbalance and applied an ensemble technique in the classification tasks. The performances of our submissions are above the median F_1-scores in all tasks except for Task 1a. Furthermore, our model achieved first place in Task 4 and obtained a 7% higher F_1-score than the median in Task 1b.

2 Classification Tasks

2.1 Problem Definition and Datasets

We participated in six classification tasks including Task 1a: Classification of adverse effect mentions in English tweets; Task 3a and 3b: Classification of change in medication regimen in tweets and drug reviews from WebMD.com; Task 4: Classification of tweets self-reporting adverse pregnancy outcomes; Task 5: Classification of tweets self-reporting potential cases of COVID-19; and Task 6: Classification of COVID19 tweets containing symptoms. Further details about the data can be found in Magge et al. (2021). Among the six classification tasks, Task 6 was three-way classification and used micro-averaged F_1-score as the evaluation metric, while the remaining tasks were binary classification and used the F_1-score for the positive class for evaluation. For all tasks, we split the training data into a training set (`TRN`) and a validation set (`TRN_VAL`) with a 90/10 ratio, and evaluated the model on the validation set (`VAL`) released by the organizers.

2.2 Method

We used a uniform framework for all classification tasks, which consists of a Transformer-based en-

Proceedings of the Sixth Social Media Mining for Health Workshop 2021, pages 52–57

Task	Task 1a	Task 3a	Task 3b	Task 4	Task 5	Task 6	Task 1a$_o$	Task 3a$_o$	Task 5$_o$
BT	64.5	59.6	88.4	89.3	71.6	98.4	67.2	60.4	71.9
CL	62.4	55.3	87.0	83.3	67.2	98.0	63.6	54.2	66.4
RBB	71.9	57.6	89.0	89.4	74.9	98.2	75.4	60.3	75.8
RBL	68.4	61.4	88.8	92.0	76.5	98.6	**78.6**	62.4	76.8
RBB+BT	66.7	62.5	89.1	91.2	79.2	98.6	73.9	64.7	77.0
RBB+RBL	69.1	66.7	89.4	92.7	80.3	98.8	75.2	65.6	79.2
RBB+CL	66.7	59.1	89.1	88.2	75.4	98.4	69.6	62.4	74.7
BT+RBL	68.5	65.7	89.5	**92.9**	78.7	**99.0**	76.5	**66.9**	78.3
BT+CL	66.7	59.4	88.7	87.4	72.8	98.4	67.9	61.7	74.1
RBL+CL	65.4	60.0	89.3	90.6	74.6	98.6	73.7	63.2	74.4
RBB+BT+RBL	67.3	64.9	89.4	92.5	**80.8**	98.8	74.3	66.4	79.1
RBB+BT+CL	67.3	61.5	89.0	89.7	74.8	98.8	67.9	63.8	76.7
RBB+RBL+CL	68.5	61.4	89.7	91.4	76.7	98.6	73.7	66.1	77.5
BT+RBL+CL	66.7	61.2	89.5	91.4	76.4	**99.0**	71.6	65.3	77.2
BT+CL+RBB+RBL	66.7	61.9	**89.8**	91.8	78.2	98.6	75.0	65.6	79.1

Table 1: F_1-scores of individual models and ensemble models on the validation (VAL) sets, where **Task***$_o$ denotes that the models are trained on the oversampled training (TRN) sets, and **ALL** denotes the ensemble of four individual models. The model that performed best on each task is highlighted in boldface.

coder, a pooling layer, a linear layer, and an output layer with Softmax activation. For each instance, the encoder converts each token into an embedding vector, and the pooling layer generates a document embedding by averaging the token embeddings. The document embedding is then fed into the linear layer and the output layer. The output is a probability vector with values 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 during the inference phase.

Encoder: Encouraged by the success of pre-trained Transformer-based language models, we experimented on four Transformer-based models pre-trained on different corpora–BERTweet (BT) (Nguyen et al., 2020) trained on English tweets, Bio_Clinical BERT (CL) (Alsentzer et al., 2019) on biomedical research papers and clinical notes, and RoBERTa$_{Base}$ (RBB) and RoBERTa$_{Large}$ (RBL) (Liu et al., 2019) on generic text such as English Wikipedia. We selected these models in order to investigate how the model size and the domain of pre-training data can benefit the performance on health-related tasks with social media data.

Preprocessing: To reduce the noise of tweets, we used the open source tool `preprocess-twitter` for data preprocessing.[1] The preprocessing includes lowercasing, normalization of numbers, usernames, urls, hashtags and text smileys, and adding extra marks for capital words, hashtags and repeated letters.

Oversampling: As described in Magge et al. (2021), the class distributions of Task 1a, Task 3a and Task 5 are imbalanced. To address the problem, we oversampled the minority class in the training set by picking samples at random with replacement using a Python toolkit called *imbalanced-learn*. The script is available on Github.[2] After oversampling, the new training sets included 28,942, 9644 and 9786 instances for Task 1a, Task 3a and Task 5, respectively.

Ensemble Modeling: In an attempt to improve performance over individual classifiers, we applied an ensemble technique to combine the results of different models. We averaged the outputs (*i.e.*, the probability vectors) of each model and selected the class with the highest value as the inference result.

2.3 Experiments and Results

We trained each model for 10 epochs, and the checkpoints that achieved the best performances on TRN_VAL were selected for evaluation. We experimented with two learning rates $\in \{2^{e-5}, 3^{e-5}\}$ and three different random initializations, meaning that there were six checkpoints in total for each model.[3] For each type of model, the median of the six checkpoints was used when we reported the results of individual models (*i.e.*, BT, CL, RBB, and RBL). For each ensemble model, all of the six checkpoints of the same type of model were

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

[2] https://gist.github.com/yguo0102/c72b5c0c353bea31bd7d72a15f6a0899

[3] Other hyper-parameters are fixed for all models. The batch size is 32, the weight decay is 0.1, and the warm-up ratio is 0.06.

used. Therefore, an ensemble model that combines k types of models consists of $6 \times k$ checkpoints.

	Task	Precision	Recall	F_1-Score
Ours	Task 1a	52.1	32.7	40.0
	Task 3a	72.1	63.5	68.0
	Task 3b	84.2	88.2	86.0
	Task 4	**93.9**	**92.2**	**93.0**
	Task 5	73.2	77.3	75.0
	Task 6	94.5	94.5	94.0
Median	Task 1a	50.5	40.9	44.0
	Task 4	91.8	92.3	92.5
	Task 5	73.9	74.4	74.5
	Task 6	93.2	93.2	93.0

Table 2: Our results and the median results on the evaluation sets of the classification tasks. The system that ranked first during the competition is highlighted in boldface.

Table 1 shows the results of individual models and ensemble models trained on the oversampled training sets. For each task, we submitted the model that performed best on the validation set, and the results of the test sets are shown in Table 2. The performances of our systems were above the median for each task except for Task 1a, and achieved first place on Task 4. For Task 3a and Task 3b, our system achieved 18% higher F_1-score on Task 3a and comparable result on Task 3b compared to the baseline model (Weissenbacher et al., 2020).[4]

2.4 Analysis

In general, for individual models, RoBERTa$_{Base}$ and RoBERTa$_{Large}$ performed better or comparable to BERTweet, and Bio_Clinical BERT underperformed on all tasks compared to the other models, which is consistent with our previous findings (Guo et al., 2020). Ensemble models outperformed individual models on all tasks except for Task 1a. We observed that for Task 1a, all models achieved high F_1-scores (around 97%) on the `TRN_VAL` set after training for 1 epoch, but the performance dropped by 25%-35% on the `VAL` set. Similarly, our F_1-score on the testing set of Task 1a is 40%, which is lower than that on the `VAL` set. Since the same trend is not present for other tasks, we hypothesized that the types of ADE in the training set and validation set of Task 1a may have low overlap.

To test our hypothesis, we counted the number of distinct ADE labels and normalized ADE labels using the data of Task 1b and Task 1c, shown in Table 3. Interestingly, the overlap percentage of normalized ADE labels is as high as 85.5%, and that of unnormalized ADE labels is much lower. This suggests that most types of ADE in the validation set are included in the training set but the ADE descriptions can vary widely. This result indicates that the gap between the performance on the training set and validation set may be attributed to the limited generalizability of pre-trained Transformer-based models to capture the semantic similarities between different expressions of the same ADE.

Type	Training	Validation	Overlap/percent
ADE	1127	85	35/41.2%
ADE_n	476	69	59/85.5%

Table 3: The number of the distinct ADE labels in the training set and validation set of Task 1, where **ADE**$_n$ denotes the normalized ADE labels. The overlap percentage is computed based on the validation set.

3 Task 1b - ADE Span Detection

3.1 Problem Definition and Dataset

Task 1b aims at distinguishing adverse effect mentions from Non-ADE expressions and identifying the text spans of these adverse effect mentions. A tweet can have more than one ADE mention, and an ADE mention can be a sequence of words as well. The training set consists of 17,385 tweets annotated with 1713 ADE mentions for 1235 tweets, and the validation set consists of 915 tweets annotated with 87 ADE mentions for 65 tweets.

3.2 Method

We implemented several Transformer-based models including BioBERT (Lee et al., 2020), SciBERT (Beltagy et al., 2019), BERTweet (Nguyen et al., 2020) and two models of BERT (Devlin et al., 2019), and compared their performances.[5] BioBERT is specifically trained for biomedical text and widely used for the biomedical text-mining for NER. SciBERT is trained on more general domain data such as computer science text. BERTweet is a pre-trained language model for English Tweets. In addition, since the dataset is very imbalanced, we also performed undersampling to change the composition of the training set. Specifically, we randomly divided the training data with negative labels into 10 non-overlapping subsets, each of which

[4] Because we were the only participant for Task 3a and 3b, there is no median score available.

[5] For each of these 5 methods, we used the "cased" and "base" models if it is not specified.

has a slightly larger size (2000 tweets) compared to the positive data (the same 1235 positive tweets), and then 5 subsets were randomly selected for our experiment.

3.3 Experiments and Results

In our experiments, since the tweets are relatively short, we set the max sequence length to 128, batch size to 128 for BERT$_{Large}$ and 256 for other models. The learning rate was set to 5^{e-5}, and the epoch was set to 20 for all 5 models. The final submissions were evaluated in terms of precision, recall, and F_1-score by the official evaluation scripts provided by the organizers, for each ADE extracted where the spans overlap either entirely or partially. However, for the convenience of comparing the performance of the models during the experiments, we used Seqeval,[6] which is a Python framework for sequence labeling evaluation, to compare all methods on the validation set also by precision, recall, and F_1-score at the token level. Table 4 shows the performances for these 5 models.

Model	Precision	Recall	F_1-score
BioBERT	38.0	42.2	40.0
BERTweet	36.6	41.3	38.8
SciBERT	44.3	42.2	43.2
BERT$_{Base}$	**48.1**	39.1	43.1
BERT$_{Large}$	47.6	**46.9**	**47.3**

Table 4: The performances of models on validation set. The highest scores of precision, recall, and F_1-score have been highlighted in the table respectively.

From Table 4, it can be observed that BERT$_{Large}$ outperforms all other models with the highest recall and F_1-score. As a result, we chose BERT$_{Large}$ as the model used in the final submission. Finally, the result we received from the organizers was similar to the performance on the validation set, which is above the median. Although our recall is 17% worse than the median recall, our precision is 68.1 (+19%) and our F_1-score is 49.0 which is 7% higher than the median F_1-score.

3.4 Analysis

3.4.1 Comparison Between Models

We conducted the research on the learning efficiency and the performance over 20 epochs of each model, evaluating each time on the validation set. The results of precision, recall, and F_1-score for each epoch are shown in Figure 1.

These three plots show that the learning efficiency of BERT$_{Large}$ is very fast. When the epoch is 2, precision, recall and F_1-score for this model reach about 35%, while the scores of other models are only around 15% at this stage. In addition, as shown in the plots, the performance of BERT$_{Large}$ is consistently better than other models during training, which may benefit from its larger pre-training dataset. However, it is surprising to find that, unlike the curves of BioBERT, SciBERT and BERTweet, the curves of BERT$_{Base}$ model are relatively unstable, with some fluctuations.

3.4.2 Undersampling Experiments

Since BERT$_{Large}$ was the best model in our experiments, we separately finetuned BERT$_{Large}$ for 10 epochs on each of the 5 undersampled datasets, and compared the average scores for these 5 subsets with the performance scores obtained without undersampling. These results were also evaluated on the validation set at the token level. The results for undersampling are shown in Table 5. The averaged F_1-score for all the undersampled subsets is significantly lower than the best performance. Although we used all the positive data, it is possible that the drastic reduction in the amount of negative data and the total training data has had a very large impact on the results. Furthermore, randomly sampling the negative examples changes the prior distribution of the probability for the classifier. Due to time constraints associated with the shared task deadline, we were unable to try more advanced heuristics to select the negative examples for the undersampling, which is worth further exploring in future work.

Model	Precision	Recall	F_1-score
Subset-data1	22.1	62.5	32.7
Subset-data2	24.7	60.9	35.1
Subset-data3	22.7	63.6	33.3
Subset-data4	25.4	68.8	37.1
Subset-data5	23.0	64.1	33.9
AVG of Subset-data	**23.6**	**64.0**	**34.4**
With all training data	**47.6**	**46.9**	**47.3**

Table 5: Results when training on 5 undersampled datasets. Subset-data1 to Subset-data5 represent the 5 subsets that were randomly selected for experiment.

3.4.3 Performance Analysis

In order to conduct the research on the results we received from the organizers, we compared the annotated data for validation set provided by the or-

[6] https://github.com/chakki-works/seqeval

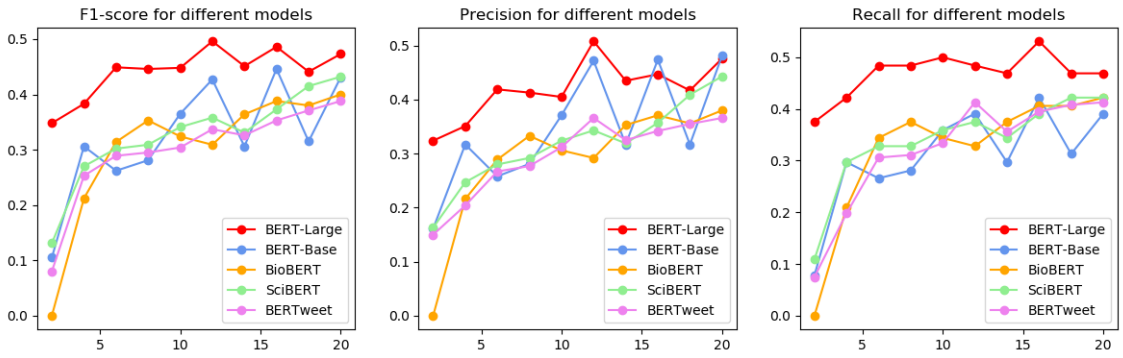

Figure 1: F_1-score, Precision and Recall for different models applied to the span detection task. The x-axis represents epoch.

ganizers with the results predicted by BERT_{Large}. This analysis revealed two primary causes why our model did not receive higher scores. Firstly, the number of true positives is relatively small. 87 annotations with label "ADE" were given in the validation set, but after the prediction, only 60 ADE mentions in the validation set (including true positive cases and false positive cases) were obtained. In these 60 ADE mentions, 23 cases which we only partially correctly predicted are also included, which means that many true ADEs were not detected (false negatives). Secondly, most of the ADE mentions predicted by our models, which are not annotated with label "ADE" in the validation set, did not appear for no reason, but actually have been annotated with label "ADE" in the training set. For example, "nosleep", which does not seem to have any ambiguity, is marked as ADE in one tweet, but not in another tweet, which might be due to the differences in the contexts in which they are mentioned. For example, in some tweets, "nosleep" appears in the tag "teamnosleep"; although it was predicted as ADE mention after being tokenized, it was not actually labeled as ADE by annotators.

4 Conclusion

In this work, we developed two systems based on pre-trained Transformer-based models for multiple health-related classification tasks and one span detection task for the SMM4H 2021 shared tasks. We experimented with different Transformer-based model variants as well as sampling strategies and applied an ensemble technique in the classification tasks. The results of our submissions are over the median F_1-scores in all tasks except for Task 1a. Furthermore, our model achieved first place in Task 4 and obtained a 7% higher F_1-score than the median in Task 1b. For future work, we will investigate methods to improve the generalizability of pre-trained Transformer-based models to deal with various health-related expressions in social media data.

References

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.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. volume 1903.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. volume 1810.

Yuting Guo, Xiangjue Dong, Mohammed Ali Al-Garadi, Abeed Sarker, Cecile Paris, and Diego Mollá Aliod. 2020. Benchmarking of Transformer-Based Pre-Trained Models on Social Media Text Classification Datasets. In *Proceedings of the The 18th Annual Workshop of the Australasian Language Technology Association*, pages 86–91, Virtual Workshop. Australasian Language Technology Association.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. In *Bioinformatics*, volume 36, page 985–989.

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 Klein, Ivan Flores, Ilseyar Alimova, Mohammed Ali Al-garadi, Antonio Miranda-Escalada, Zulfat Miftahutdinov, Eulàlia Farré-Maduell, Salvador Lima López, Juan M Banda, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Martin Krallinger, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021. Overview of the sixth social media mining for health applications (# SMM4H) shared tasks at NAACL 2021. In *Proceedings of the Sixth Social Media Mining for Health Applications Workshop & Shared Task*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. 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.

Davy Weissenbacher, Suyu Ge, Ari Klein, Karen O'Connor, Robert Gross, Sean Hennessy, and Graciela Gonzalez-Hernandez. 2020. Active Neural Networks to Detect Mentions of Changes to Medication Treatment in Social Media. *medRxiv*.