Automatic Detecting for Health-related Twitter Data with BioBERT

Yang Bai, Xiaobing Zhou*
School of Information Science and Engineering
Yunnan University, Yunnan, P.R. China
⋆Corresponding author:zhouxb@ynu.edu.com

Abstract

Social media used for health applications usually contains a large amount of data posted by users, which brings various challenges to NLP, such as spoken language, spelling errors, novel/creative phrases, etc. In this paper, we describe our system submitted to SMM4H 2020: Social Media Mining for Health Applications Shared Task which consists of five sub-tasks(Ari Z. Klein and Gonzalez-Hernandez., 2020). We participate in subtask 1, subtask 2-English, and subtask 5. Our final submitted approach is an ensemble of various fine-tuned transformer-based models. We illustrate that these approaches perform well in imbalanced datasets (For example, the class ratio is 1:10 in subtask 2), but our model performance is not good in extremely imbalanced datasets (For example, the class ratio is 1:400 in subtask 1). Finally, in subtask 1, our result is lower than the average score, in subtask 2-English, our result is higher than the average score, and in subtask 5, our result achieves the highest score.

1 Introduction

According to the United States Centers for Disease Control and Prevention (CDC), drugs administered for alleviating common sufferings are the fourth biggest cause of death and birth defects are the leading cause of infant mortality. These are the most important medical problems for human society.(Giacomini et al., 2007)

Twitter, a popular micro-blogging service, has received much attention recently. It is an online network used by millions of people around the world to stay connected to their friends, family members, and co-workers through their computers and mobile telephones(Barbosa and Feng, 2010). On average, one in a thousand messages from public Twitter data is health-related. These health-related Twitter posts can help us analyze various human health-related phenomena. For example, there are limited methods for studying birth defects in infants, and this knowledge has been challenging(Klein et al., 2018)(Klein et al., 2019). This situation provides a challenging opportunity due to the increasing number of related tweets on Twitter.

With this motivation, five shared tasks are conducted as part of the Social Media Mining for Health Applications (SMM4H) Workshop 2020 hosted by the University of Pennsylvania Health Language Processing (HLP) Lab. Our team participated in subtasks 1, 2-English, and 5 of the workshop. Some samples from the training set are given in Table 1, and these tasks are:

- Sub-task 1: Automatic classification of tweets that mention medications

  We take subtask-1 as a binary classification task. The model is required to determine whether a medication or dietary supplement is mentioned in a tweet, and predict a label L, where $L \in \{$mention a medication or dietary supplement - 1, no mention - 0$\}$.

- Sub-task 2-English: Automatic classification of English tweets that report adverse effects

Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 63–69
Barcelona, Spain (Online), December 12, 2020.

We take subtask-2-English as a binary classification task. The model is required to determine whether an adverse effect of medication is reported in a tweet, and predict a label L, where $L \in \{$ report adverse effects of medication - 1, no report - 0$\}$.

- Sub-task 5: Automatic classification of tweets reporting a birth defect pregnancy outcome

  We take subtask-5 as a ternary classification task. The model is required to determine whether the user has a child and indicate that the child has the birth defect mentioned in a tweet, and predict a label L, where $L \in \{$ refer to the users child and indicate that he/she has the birth defect mentioned in the tweet - 1, ambiguous about whether someone is the users child and/or has the birth defect mentioned in the tweet - 2, merely mention birth defects - 3$\}$.

Sub-task	Sentence	Gold Label
Sub-task 1	@username you can try vitamins for Olivia! Hudson has taken vitamins since the NICU &amp; it helps him gain weight	1
	Can someone get me some chippy chips please	0
Sub-task 2 english	today i'm an emotional mess. that's what happens when i think i could wean myself off of cymbalta	1
	@username. prozac makes us all better people. #prozacnation	0
Sub-task 5	@uesername, my baby was born with microcephaly in the US due to #cytomegalovirus not #zika. Why no warning about that? #CDCchat"	1
	"Username @Usrename in a way, mine are advantaged by the older having Down's syndrome. I can recommend it!"	2
	Girl born without an ear can finally hear properly for the first time after surgeons rebuild it	3

Table 1: Samples from the training set of three subtasks

We try a variety of approaches on these tasks, including classical machine learning methods, CNN models, RNN models, and transformer-based models. We find that the transformer-based neural models consistently outperform other methods. The framework of our implementation is shown in Figure 1. Firstly, we combine the official training set and the validation set to get the new data set, which is split into the new training set and the validation set by using the stratified 5-fold cross-validation[1]. Secondly, the test set is predicted by fine tuning the model. Thirdly, we create pseudo-label to combine training set and input these data into model training and prediction. Finally, we get the final result by hard voting.

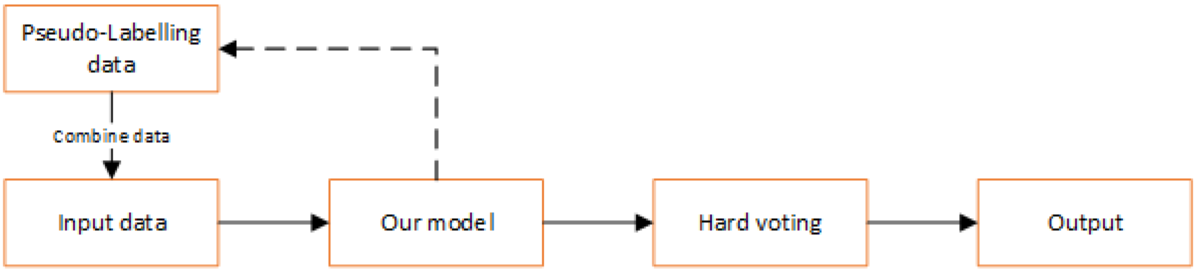


Figure 1: Our framework

2 Related Work

Normally, data in the health field is difficult to obtain. Fortunately, the existence of social media such as Twitter has alleviated this situation. However, it has caused problems that the data is difficult to handle.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html

Because diseases and other health problems belong to a small number of people. This phenomenon has caused data imbalance in social media. These problems bring great challenges.

Much previous work has focused on tracking and monitoring diseases on social media. (Yin et al., 2015) has developed a scalable system by training classifiers on a dataset of 34 health topics. It created a general health classifier using standard SVM. Recently, the use of dense word vectors or embeddings is becoming popular, while previous models (e.g. Word2Vec (Pennington et al., 2014), GloVe (Pennington et al., 2014), FastText (Joulin et al., 2016)) focus on learning context-independent word representations, recent works have focused on learning context-dependent word representations. For instance, BERT (Devlin et al., 2018) is a contextualized word representation model, which is based on a masked language model and pre-trained by using bidirectional transformers(Weissenbacher et al., 2018; Weissenbacher et al., 2019b). BioBERT is a domain-specific language representation model pre-trained on a large biomedical corpus(Lee et al., 2020). In this paper, we use the BioBERT in our experiments and use it for the different text classification tasks of Social Media Mining for Health Workshop. In the following, we describe our dataset and the methods used for different tasks.

3 Methodology

3.1 Dataset

The data sets of these tasks are all from social media. These data sets contains rare health-related events such as pregnant women groups, drug effects, birth defects, etc(Sarker et al., 2017; Weissenbacher et al., 2019a; O'Connor et al., 2020).

- For shared task 1. In the training set, 55,419 tweets are included, with 146 tweets mentioning medications (1) and 55,273 tweets not mentioning (0). In the validation set, 13853 tweets are included, 35 labeled 1, and 13818 tweets labeled 0. This is an extremely imbalanced data set, so the evaluation indicator for this task is F1-score for the positive class (i.e., tweets that mention medications).

- For shared task 2. In the training set, 20,544 tweets are included, 1903 tweets that report adverse effects of medications (1), and 18461 tweets that do not report (0). In the validation set, 5134 tweets are included, 474 tweets that report adverse effects of medications (1), and 4660 tweets that do not report (0). This is an imbalanced data set, so the evaluation indicator for this task is F1-score for the positive class (i.e., tweets that report adverse effects of medications).

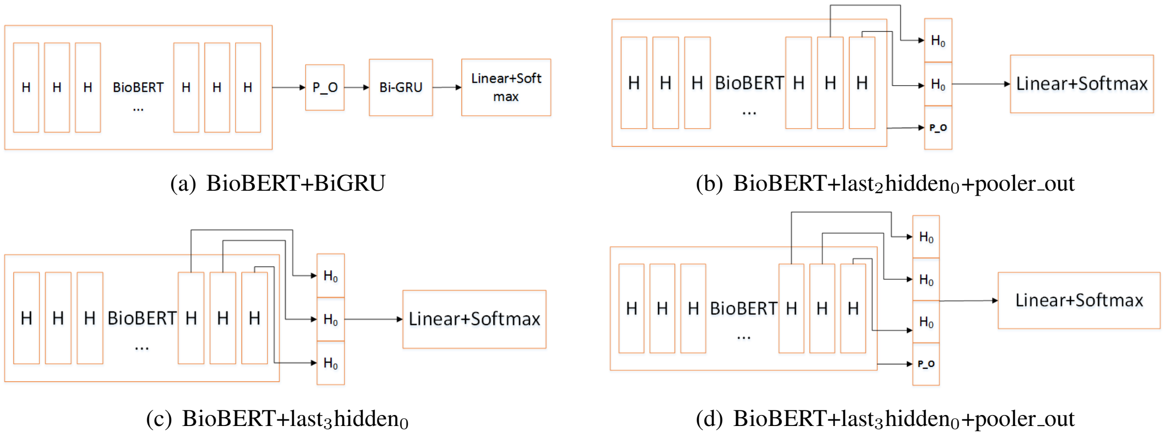
- For shared task 5. In the training set, there are 14717 tweets, 773 tweets for 'defect' (1), 834 tweets for 'possible defect' class (2), and 13110 tweets for 'non-defect' (2). In the validation set, there are 3680 tweets, 193 tweets for 'defect' (1), 207 tweets for 'possible defect' (2), and 3280 tweets for 'non-defect' (3). This is an imbalanced data set, so the evaluation indicator for this task is micro-averaged F1-score for the "defect" and "possible defect" classes.

3.2 Models

There are some limitations in applying NLP directly to biomedical text mining. With the recent word representation model (such as word2vec), Elmo and BERT are trained and tested on data sets containing common domain text. However, these models do not perform well on biomedical text data sets. So, we choose the BioBERT[2] as our model for these tasks we participated in. For classification tasks, the output of BioBERT(pooler output) is obtained by its last layer hidden state of the first token of the sequence (CLS token) further processed by a linear layer and a tank activation function. But the pooler output is usually not a good summary of the semantic content of the input. So we try the following model architecture to relieve this problem. Our model architecture is shown in Figure 2.

For Figure 2 (a), we can regard the model as two parts. The first part is to get the output of BioBERT(P_O). The second part is the BiGRU module with input P_O. For Figure 2(b), we concatenate P_O and H_0 of the last two hidden layers into the classifier after obtaining P_O. For Figure

[2] https://github.com/vthost/biobert-pretrained-pytorch/releases/tag/v1.1-pubmed



(a) BioBERT+BiGRU

(b) BioBERT+last$_2$hidden$_0$+pooler_out

(c) BioBERT+last$_3$hidden$_0$

(d) BioBERT+last$_3$hidden$_0$+pooler_out

Figure 2: The models we used (H is the hidden layer of BioBERT, P_O is the pooler output, H_0 is hidden-state of the first token of the sequence(CLS token) at the output of the hidden layer of the model.)

2(c), we concatenate H_0 of the last two hidden layers into the classifier. For Figure 2(d), we concatenate P_O and H_0 of the last three hidden layers into the classifier after obtaining P_O.

3.3 Pseudo Label

In order to make our model more robust, we use pseudo labels. First, we input the training set and the validation set into the $modelA$ for training, and predict a $resultA$ in the test set. Secondly, we add 10% of $resultA$ to the training set to obtain a training set with pseudo-label. Again, we input the training with pseudo-label and validation set into the $modelA$ for training, and predict a final result in the test set. The process of pseudo label is shown in Figure 3.

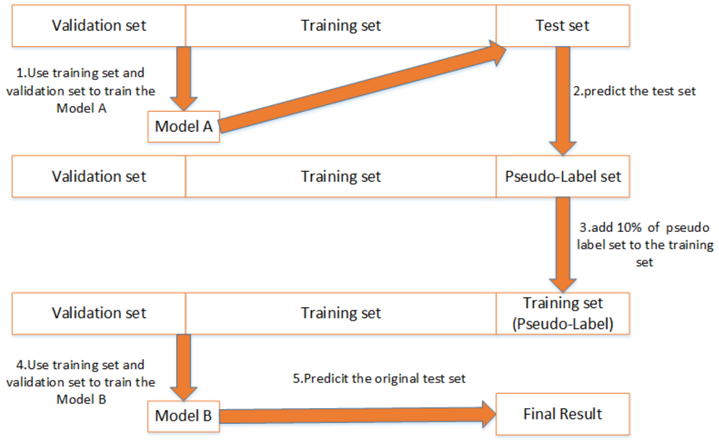

Validation set
Training set
Test set
1.Use training set and validation set to train the Model A
Model A
2.predict the test set
Validation set
Training set
Pseudo-Label set
3.add 10% of pseudo label set to the training set
Validation set
Training set
Training set (Pseudo-Label)
4.Use training set and validation set to train the Model B
5.Predicit the original test set
Model B
Final Result

Figure 3: The process of pseudo label

4 Experiments Result

4.1 Experiments Setup

BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) is a domain-specific language representation model pre-trained on large-scale biomedical corpora. It greatly outperforms BERT and previous advanced models in many biomedical text mining tasks. For all tasks, we use the BioBERT-Base-v1.1-PubMed model and RAdam(Liu et al., 2019) as BioBERT optimizer. In order to increase the difference in model fusion, we preprocess the data input by some models. Table 2 shows the hyper-parameters used by our different models. Preprocessing data is simple data cleaning of

	Models	Hyperparameters

1	5-fold data with 42 random seeds
	BioBERT+BiGRU
		output_hidden_states=False
		dropout=0.1
		learning rate=3e-5
		epoch=3
		per_gpu_train_batch_size=4
		gradient_accumulation_steps=4

2	5-fold data with 42 random seeds
	BioBERT+last$_2$hidden$_0$+pooler_out
		output_hidden_states=True
		dropout=0.1
		learning rate=3e-5
		epoch=3
		per_gpu_train_batch_size=4
		gradient_accumulation_steps=4

3	5-fold data with42 random seeds BioBERT+last$_3$hidden$_0$	
4	5-fold data with 24 random seeds BioBERT+BiGRU	
5	5-fold data with 24 random seeds BioBERT+last$_2$hidden$_0$+pooler_out	
6	5-fold data with 42 random seeds Preprocess data BioBERT+last$_2$hidden$_0$+pooler_out	
7	5-fold data with 42 random seeds Pseudo Label Preprocess data BioBERT+BiGRU	
8	5-fold data with 42 random seeds Pseudo Label Preprocess data BioBERT+last$_2$hidden$_0$+pooler_out	

9	5-fold data with 42 random seeds
	BioBERT+last$_3$hidden$_0$+pooler_out
		output_hidden_states=True
		dropout=0.1
		learning rate=2e-5
		epoch=3
		per_gpu_train_batch_size=4
		gradient_accumulation_steps=4

Table 2: Hyperparameters of these models in our experiments for three subtasks.

the data set such as: removing URL, standardizing case, and standardizing @ username. The purpose of preprocessing data is to increase the difference of results during model ensemble.

From table 2, we can see the Hyperparameters of these models. Firstly, we combine the official training set and the verification set to get the new data set, which is split into the new training set and the verification set by using the stratified 5-fold cross-validation. Secondly, we input data into one to six models for training with the training set and predict six results with the test set. We combine these six results($result1 - 6$) to get $result - A$ by hard voting. Thirdly, in order to make pseudo label, we take 10% of the data set from $result - A$ which is combined into the training set. We input the data set with pseudo-label into the seven-eight model to train and predict, we get these result ($result7 - 8$). Fourthly, we input the data into the ninth model training and prediction, we get the result($result9$). Finally, we combine $result1 - 9$ by hard voting to get the final result.

4.2 Results

Table 3 presents the performance scores for three subtasks on the test data, our results are based on hard voting of the 9 models in Table 2. For all tasks, we only use the official training set and validation set and do not use any external data. As can be seen from the table, in extremely imbalanced data, our method

		Precision	Recall	F-1 Score
Subtask-1	Our score	0.8649	0.4156	0.5614
	Mean score	0.7007	0.7039	0.6646
Subtask-2 english	Our score	0.49	0.60	0.54
	Mean score	0.42	0.59	0.46
Subtask-5	Our score	0.65	0.73	0.69
	Mean score	0.62	0.68	0.65

Table 3: The results of our method on the test set for three subtasks

has no advantage. But in imbalanced data, our method has achieved good results.

5 Conclusion

In this work, our method is an ensemble of various fine-tuned transformer-based models on these tasks. We obtain decent results for these tasks organized as a shared task in Social Media Mining for Health Workshop - 2020. Our biggest regret in this work is that in extremely imbalanced data sets, we have not done too much processing on the data sets and can't achieve promising results. In the future, our work will focus on solving the problem of extremely imbalanced data. The code is available online.[3]

References

Ivan Flores Arjun Magge Zulfat Miftahutdinov Anne-Lyse Minard Karen O'Connor Abeed Sarker Elena Tutubalina Davy Weissenbacher Ari Z. Klein, Ilseyar Alimova and Graciela Gonzalez-Hernandez. 2020. Overview of the fifth social media mining for health applications (#SMM4H) shared tasks at COLING 2020. in proceedings of the fifth social media mining for health applications (#SMM4H) Workshop & Shared Task.

Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on Twitter from biased and noisy data. In *Coling 2010: Posters*, pages 36–44.

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*.

Kathleen M Giacomini, Ronald M Krauss, Dan M Roden, Michel Eichelbaum, Michael R Hayden, and Yusuke Nakamura. 2007. When good drugs go bad. *Nature*, 446(7139):975–977.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

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

Ari Z Klein, Abeed Sarker, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2019. Towards scaling Twitter for digital epidemiology of birth defects. *NPJ digital medicine*, 2(1):1–9.

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. *Bioinformatics*, 36(4):1234–1240.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2019. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*.

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

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

[3] https://github.com/byew/SMM4H-2020

Abeed Sarker, Pramod Chandrashekar, Arjun Magge, Haitao Cai, Ari Klein, and Graciela Gonzalez. 2017. Discovering cohorts of pregnant women from social media for safety surveillance and analysis. *Journal of medical Internet research*, 19(10):e361.

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

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

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen OConnor, Michael Paul, and Graciela Gonzalez. 2019b. Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019. In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, pages 21–30.

Zhijun Yin, Daniel Fabbri, S Trent Rosenbloom, and Bradley Malin. 2015. A scalable framework to detect personal health mentions on Twitter. *Journal of medical Internet research*, 17(6):e138.