Automatic Classification of English Tweets Reporting Adverse Effects of Medication by Deep Learning

Lingyao Meng and Richard Wang

School of Information, University of California, Berkeley, CA, USA

{lingyaomeng, cweiwang}@berkeley.edu

Abstract

With the rapid growth of social networking users worldwide, social media has become an important source of health related information. Advances in automatic data processing and NLP provide opportunities for health and medical data mining from this massive data source. Here is presented the effort to automatically classify English tweets reporting adverse effects of medication, which is a subtask of the SMM4H'20 shared task. Previous work has shown outperformance of BERT on this task, compared to traditional machine learning systems or other deep neural architectures. We applied BERTweet, a recently released large scale pre-trained language model for English tweets on this task and obtained significantly better results than the baseline from the BERT model. Medications mentioning reddit posts were obtained and labeled as additional dataset for model training but no further improvement of model performance was observed. Future work could be sprawling medication mentioning tweets to pre-train the BERTweet model.

1. Introduction

Adverse Effect (AE) is an injury caused by taking medication. Both the pharmaceutical companies and the regulatory agencies closely monitor AEs because it could result in serious safety issues. With the rapid growth of social media users worldwide, online posts, especially tweets become an important source of information of AEs. Under the current situation of COVID-19, more twitter users tend to follow health related topics and share their experiences on the platform. Therefore, we foresee the uprise of AE reporting when COVID-19 therapeutics and/or vaccine becomes available. The recent progress in natural language processing (NLP) makes it possible to automatically classify whether AE is mentioned in a tweet (Weissenbacher et al., 2019), and such a procedure is necessary to select AE mentioning tweets for information extraction.

The work reported here is for subtask 2 of the SMM4H'20 shared tasks, which involve NLP challenges on social media mining for health monitoring and surveillance. This binary classification task requires systems to process in-balanced and noisy real world language expressions, take into account subtle linguistic variations between AEs and indications and predict whether AE is mentioned in a tweet or not. Previous works have shown outperformance of Bidirectional Encoder Representations from Transformers (BERT) on this task, compared to traditional machine learning systems such as SVM or CRF and other deep neural architectures such as CNN or RNN (Chen et al., 2019; Miftahutdinov et al., 2019; Ellendorff et al., 2019; Bagherzadeh et al., 2019; Anand et al., 2019). Here we took the training corporus to fine tune the BERT model and deployed the model to get the baseline predictions.

The pre-training corpora of BERT include books and Wikipedia, while tweets usually exhibit different characteristics from such resources (Devlin et al., 2018). The frequent use of informal grammar and irregular vocabulary might lead to a challenge in applying BERT to process tweets. The recently released model BERTweet is a pre-trained BERT model for English tweets. Researchers of BERTweet trained BERT with a large scale corpus of English tweets based on the RoBERTa pre-training procedure (Liu et al., 2019) and claimed its outperformance on several tweet NLP tasks (Nguyen et al., 2020).

In this work, we fine tuned the BERTweet model with the AE classification training data and achieved significantly better results than the baseline. In the effort to further improve the model performance, we collected additional medication mentioning posts from Reddit and used them as the "silver" dataset for training.

2. Data

The organizers of the shared tasks provided the participants data for training and evaluation. The training data contains 25,678 tweets with 2,377 "positive" tweets and 23,301 "negative" tweets. The evaluation data contains 4759 tweets. The evaluation metric is the f1-score for the "positive" class (i.e., tweets that report AEs). The distribution of "positive" class within the evaluation data was not disclosed.

Two annotators with biomedical education and both experienced in Social Media research tasks manually annotated the corporus. The annotators independently dual-annotated each test set to insure the quality of annotations. Disagreements were resolved after an adjudication phase between two annotators. The inter annotator agreement (IAA) was high, with a Cohens Kappa = 0.82 (Weissenbacher et al., 2019).

3. Methods

3.1 Data pre-processing

For the purpose of data quality and dimension reduction, tweets were pre-processed to remove URLs, hashtags, mentions, numbers, emojis and smileys. It was done using a preprocessing library for tweet data (Özcan, 2020). Some training examples before and after the process are shown in Table 1.

Table 1. Selected Tweets Before and After Pre-process

Original tweets	Processed tweets
@for_esme yeah, i've already been on cymbalta and lexapro in the last year.	yeah, i've already been on cymbalta and lexapro in the last year.
someone sell me suboxone i can trade a xbox 360	someone sell me suboxone i can trade a xbox
have you taken byetta, januvia or victoza? let us know! we can help! #defectivedrugs http://t.co/evpkb2ccbl	have you taken byetta, januvia or victoza? let us know! we can help!

a medication called latuda 👏	a medication called latuda
@josiestevensmtr trazadone or seroquel should be ok - consult a dr. 😳	trazadone or seroquel should be ok - consult a dr.

3.2 Model fine tune

The deep learning framework used for fine tuning BERT-Base and BERTweet-Base is Hugging Face and the codes were loosely based on this post: Sentiment Analysis with BERT and Transformers by Hugging Face using PyTorch and Python. Training data was randomly splitted at the ratio of 80:20 to generate the training set and the evaluation set. Hyperparameters were determined based on both the previous work (Chen et al., 2019; Miftahutdinov et al., 2019; Ellendorff et al., 2019; Bagherzadeh et al., 2019; Anand et al., 2019) and our own experiments. The optimized hyperparameters for training each model are shown in Table 2.

Table 2. Hyperparameters for Fine Tuning BERT and BERTweet

	Maximum token length	Batch size	Learning rate	Training epochs
BERT-Base	120	32	2e-5	2
BERTweet-Base	120	32	2e-5	5

Specifically, the maximum token length was set 120 as the token counts for most training examples are fewer than 60 (Figure 1). The number of training epochs were determined by the best f1-score in a training history of 20 epochs (Figure 2).

Figure 1. Distribution of Token Counts of Training Examples

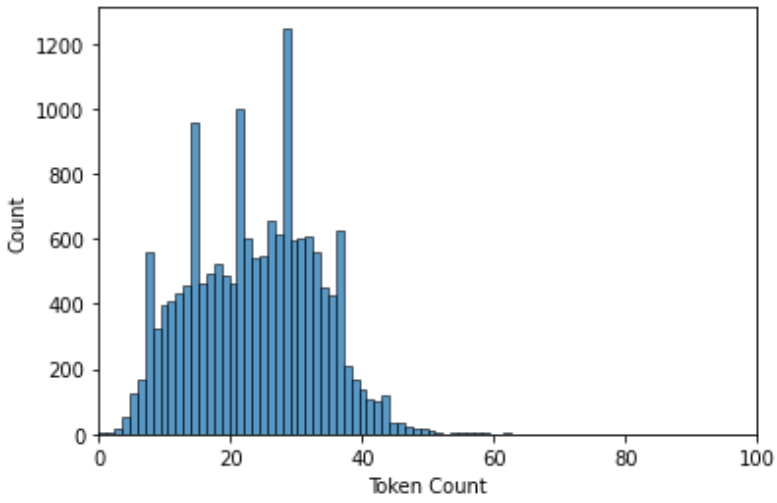
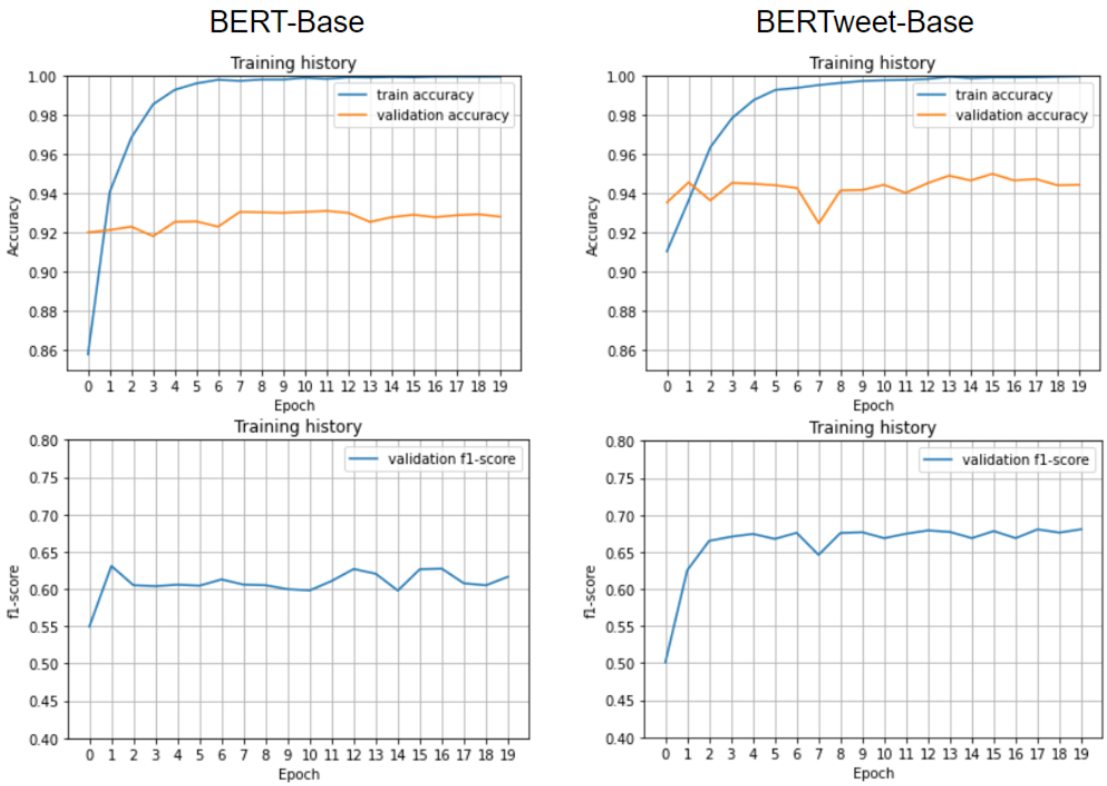


Figure 2. Training History of BERT and BERTweet for 20 Epochs



3.3 Additional dataset collection and labeling

One of the limitations we have is the size of data we have to train our model is relatively small given our model complexity. Hence, we believe collecting more data can significantly improve our model. There are three data sources available that can potentially help us including past SMM4H shared tasks, kaggle twitter sentimental competition, reddit comment dataset.

We explored past SMM4H shared tasks data as our first option. The organizers shared a list of historical tweet ids and their parsing program with us. Upon trying, the organizer's parser utilized brute force parsing method which is already banned by Twitter. We end up having to create a Twitter developer account and parser tweets through official Twitter API. We were able to obtain 2018 and 2019 shared task data. However, the data used in the 2020 shared task is identical to 2018 and 2019.

Kaggle Twitter sentiment analysis was the second option. The dataset contains over 1 million records of tweets with corresponding sentiment labelings. One of the challenges presented to us is to be able to identify tweets that are related to drug uses. In order to accomplish this, we have to come up with a list of drug names. We rank distinct words in our current dataset based on frequency and only keep the most frequent words. This provided us with a list of words that are mostly drug names with some common English words. We are not able to filter out all the common English words using NLTK's stopword corpus. We end up removing any words that appeared in the story Alice in Wonderland. This provided us with a clean list of drug names. At the end, out of 1 million tweets, roughly only a few hundreds of tweets mentioned a drug name. So we are not able to identify a meaningful size dataset.

Reddit Comment data was out last resort. The entire dataset has over 1.7 billion records of reddit comments. We were not able to process data of this size on our personal laptop. At the end, we were able to find this dataset in **Google Cloud**. We wrote SQL queries to only obtain reddit comments that directly contain a drug name. Out of a 1.7 billion record of data, we are able to obtain around 400K comments that contain a drug name. The challenge with this dataset is that we don't have the labels to train our model on. We end up downloading the Twitter Sentiment Analysis model and using the model to label our reddit dataset. The drug mentioning reddit post predicted as negative in sentiment was labeled as the "positive" class.

The BERT or BERTweet model was first trained for 1-2 epochs with the reddit dataset and then fine tuned with the training examples, using the same maximum length of tokens, batch size and learning rate shown in Table 2. The number of epochs were also determined by the best f1-score on the evaluation set in a training history of 20 epochs.

4. Results

The "positive" class precision, recall and f1-score on the evaluation set and the test set, obtained from the BERT and the BERTweet model respectively, are shown in Table 3. In fact, the BERT model provided a very strong baseline, which is consistent with the previous work (Chen et al., 2019; Miftahutdinov et al., 2019; Ellendorff et al., 2019; Bagherzadeh et al., 2019; Anand et al., 2019). As expected, the BERTweet model showed significantly better performance on both the evaluation and test set, which could be attributed to the model's better understanding and handling of linguistic characteristics of tweets.

Table 3. Model Performance on Evaluation and Test Set

	Evaluation Set			Test set		
	Precision	Recall	F1-score	Precision	Recall	F1-score
BERT-Base	0.7119	0.5527	0.6223	0.4314	0.5670	0.4900
BERTweet-Base	0.7072	0.6022	**0.6505**	0.4667	0.6495	**0.5431**

Unfortunately, no further improvement was obtained by using the additional reddit dataset. The best f1-score on the test set with silver set training on BERTweet, followed by fine tune with the golden set is 0.5270.

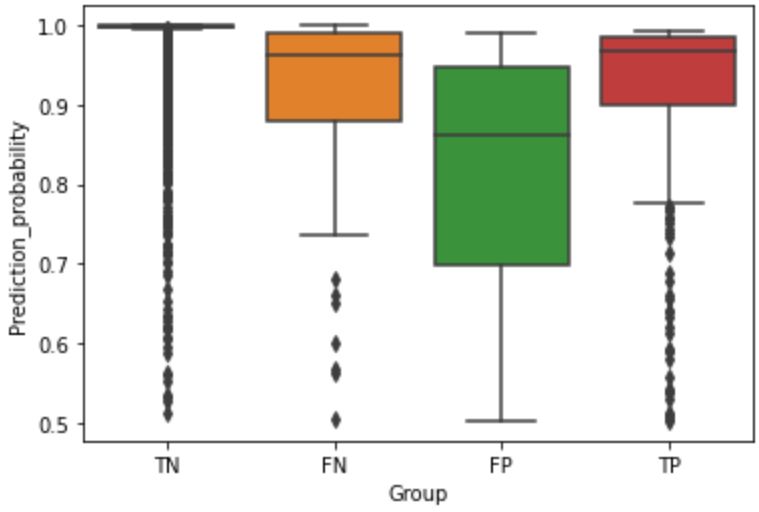
5. Discussion

Despite the superior performance of the BERTweet model, the subtle distinction between mentioning an AE and mentioning a medication's indication still cannot be comprehended very well. Similar observations were also described by previous work (Chen et al., 2019). Also, the non-trivial performance drop on the test set indicates that the fine tuned BERTweet model was not generalized enough to provide accurate predictions on new texts. Especially the low level of precision on the test set makes us believe

that the fine tuned BERTweet model tends to predict more false positives than false negatives, in percentage wise.

By investigating the evaluation set predictions, we found out that the predictions made by either the BERT or the BERTweet model, were mostly with high levels of probability. We observed only a few examples, no matter correct or wrong, predicted with probability lower than 0.8. Interestingly, the prediction probabilities of the false positive (FP) group were significantly lower than those of the true positive (TP), true negative (TN) or false negative (FN) group (Figure 3). It suggests that the model is over confident in predicting the "positive" class, and it could be the reason for the model performance drop on the test set as the "positive" class proportion is probably lower in the test set than in the training/evaluation set.

Figure 3. Prediction Probabilities on the Evaluation Set



In order to get a more generalized model, we tried to use a 20 folds large set (~400K) of medication mentioning reddit posts to train BERTweet, prior to fine tuning it with the training set. However, no further improvements were obtained from this strategy. The potential reasons are: (1) the labelings are too noisy; (2) the language styles are different between reddit posts/comments and tweets and (3) the training method is unoptimized, i.e. binary classification training may not be the best way to utilize the silver set.

6. Conclusion

The BERT model provides a strong baseline for the binary classification task of predicting tweets reporting AEs of medication. We obtained significantly better results using the BERTweet model, which is a BERT based model pre-trained with a large scale of English tweets. Additional dataset was obtained and used as the silver set but no further improvement was observed. Future work could be done such as sprawling medication mentioning tweets for pre-training the BERTweet model.

Acknowledgements

We appreciate Dr. Daniel Cer from Google, Inc. and UC Berkeley for his direction and advice on this project.

References

Davy Weissenbachery, Abeed Sarkery, Arjun Magge, Ashlynn Daughtonz, Karen O'Connory, Michael Paulz and Graciela Gonzalez-Hernandez (2019). Overview of the Fourth Social Media Mining for Health (SMM4H) Shared Task at ACL 2019. *Proceedings of the 2019 ACL Workshop SMM4H: The 4th Social Media Mining for Health Applications Workshop & Shared Task.*

Shuai Chen, Yuanhang Huang, Xiaowei Huang, Haoming Qin, Jun Yan and Buzhou Tang (2019). HITSZ-ICRC: A Report for SMM4H Shared Task 2019-Automatic Classification and Extraction of Adverse Drug Reactions in Tweets. *Proceedings of the 2019 ACL Workshop SMM4H: The 4th Social Media Mining for Health Applications Workshop & Shared Task.*

Zulfat Miftahutdinov, Ilseyar Alimova and Elena Tutubalina (2019). KFU NLP Team at SMM4H 2019 Tasks: Want to Extract Adverse Drugs Reactions from Tweets? BERT to The Rescue. *Proceedings of the 2019 ACL Workshop SMM4H: The 4th Social Media Mining for Health Applications Workshop & Shared Task.*

Tilia Ellendorff, Lenz Furrer, Nicola Colic, No¨emi Aepli and Fabio Rinaldi (2019). Approaching SMM4H with Merged Models and Multi-task Learning. *Proceedings of the 2019 ACL Workshop SMM4H: The 4th Social Media Mining for Health Applications Workshop & Shared Task.*

Parsa Bagherzadeh, Nadia Sheikh and Sabine Bergler (2019). Adverse drug effect and personalized health mentions CLaC at SMM4H 2019, Tasks 1 and 4. *Proceedings of the 2019 ACL Workshop SMM4H: The 4th Social Media Mining for Health Applications Workshop & Shared Task.*

Sarthak Anand, Debanjan Mahata, Haimin Zhang, Simra Shahid, Laiba Mehnaz, Yaman Kumar and Rajiv Ratn Shah (2019). MIDAS@SMM4H-2019: Identifying Adverse Drug Reactions and Personal Health Experience Mentions from Twitter. *Proceedings of the 2019 ACL Workshop SMM4H: The 4th Social Media Mining for Health Applications Workshop & Shared Task.*

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.*

Dat Quoc Nguyen, Thanh Vu and Anh Tuan Nguyen (2020). BERTweet: A pre-trained language model for English Tweets. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.*

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.*