Ensemble BERT for classifying medication-mentioning Tweets

Huong N. Dang, Kahyun Lee, Sam Henry, and Özlem Uzuner
George Mason University
Fairfax, VA, U.S.A.
{hdang20, klee70, shenry20, ouzuner}@gmu.edu

Abstract

Twitter is a valuable source of patient-generated data that has been used in various population health studies. The first step in many of these studies is to identify and capture Twitter messages (tweets) containing medication mentions. In this article, we describe our submission to Task 1 of the Social Media Mining for Health Applications (SMM4H) Shared Task 2020. This task challenged participants to detect tweets that mention medications or dietary supplements in a natural, highly imbalanced dataset. Our system combined a handcrafted preprocessing step with an ensemble of 20 BERT-based classifiers generated by dividing the training dataset into subsets using 10-fold cross validation and exploiting two BERT embedding models. Our system ranked first in this task, and improved the average F1 score across all participating teams by 19.07% with a precision, recall, and F1 on the test set of 83.75%, 87.01%, and 85.35% respectively.

1 Introduction

Social media platforms such as Twitter have been used in various public health studies. In these studies, detecting tweets containing health-related words such as diseases, treatments and medications is a fundamental yet difficult step. These difficulties are exacerbated by the short length and informal nature of tweets, which often contain non-standard grammar, frequent misspellings, many contractions, extensive slang, and combined symbols (emojis/emoticons) to express emotion. Despite these difficulties, unique insights can be gained from exploring Twitter data which motivates further research on this task.

Task 1 of SMM4H Shared Task 2020 (Klein et al., 2020) challenged participants to develop an automatic classification system to identify tweets mentioning medications or dietary supplements. The task was formulated as a binary classification task, in which given a set of tweets a system should predict the label for each tweet. Organizers created the UPennHLP Twitter Pregnancy Corpus (Weissenbacher et al., 2019a) and divided it into training, validation, and test sets. The training and validation datasets contained tweets labeled with either positive or negative labels indicating whether or not they contain a medication or dietary supplement mention. The test set contained tweets with labels removed, and was withheld from participants until a short evaluation period during which participants predicted labels for each tweet. Participants were evaluated based on their system's performance (F1 score for the positive class) at generating labels for the test set. The class distribution of the training, validation, and test sets was highly imbalanced. Only about 0.26% of the tweets mention medications, which is typical for this task and common in practice.

As a baseline the Shared Task organizers described the `Kusuri` classifier (Weissenbacher et al., 2019a). `Kusuri` was trained in a combination of the training and validation datasets and achieved good performance in detecting positive tweets in this test dataset at F1 score 78.79%. We sought to improve upon `Kusuri` by developing a system that combines a rule-based preprocessing component, a pre-filtering component, and an ensemble of BERT-based classifiers to predict the label of each tweet (Figure 1). In our best performing system, we used 10-fold cross validation to divide the data into 10

Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 37–41
Barcelona, Spain (Online), December 12, 2020.

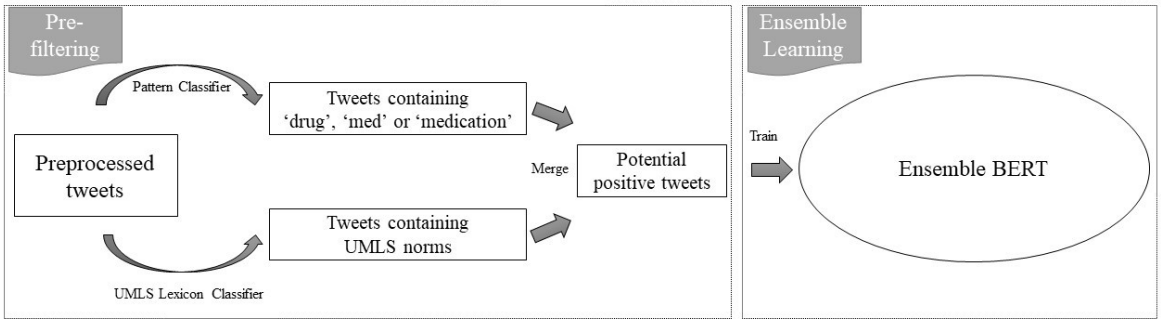

Figure 1: An overview of the system for automatic classification of tweets mentioning medications

subsets, then trained a *Bio+Clinical BERT* classifier and a *BERT-Large Uncased* classifier on each subset to obtain 20 classifiers. These 20 classifiers were ensembled using the average of their predicted probabilities to predict the final class labels. We further increased the performance of this system by training it on additional data from SMM4H 2018 which consisted of a very similar task.

2 Methods

2.1 Preprocessing

Preprocessing removes noise from tweets and converts words to a form that is recognizable by the embedding system. All tweets were pre-processed as follows: (1) Username, URL, Retweet abbreviation 'rt', non-ASCII words and characters were removed; (2) Artifacts and upstream processing such as '&', '<', '>' were standardized to 'and', '<', and '>' respectively; (3) Camel-cased expressions in hashtags were split into their component words and the hash symbols were removed; (4) A space was added between a word and a punctuation; (5) All characters were lowercased; (6) Contractions were converted to their formal forms.

2.2 Pre-filtering

Next, the preprocessed tweets were pre-filtered to remove some negative tweets. Pre-filtering was effective at removing approximately 25% of the negative tweets, and the class ratio increased from 0.26% positive class before pre-filtering to 0.34% positive class after pre-filtering. Pre-filtering consisted of a Pattern Classifier and a Unified Medical Language System Metathesaurus (UMLS) Lexicon Classifier. Tweets were labeled as *Potential positive tweets* if the requirements of either classifier were met. The Pattern Classifier labeled tweets containing the words 'drug', 'medication' or 'med' as *Potential positive tweets* based on the assumption that these tweets are likely to contain medication mentions. The UMLS Lexicon Classifier selected tweets containing at least one concept from sections 0, 1, 2 and 9 of the UMLS 2017 AA release. Given a tweet, the UMLS Lexicon classifier identified entities in that tweet and checked if those entities were linked to any concepts within a default threshold of 0.7. The UMLS Lexicon classifier was implemented using the Name Entity Recognition (NER) model and Entity Linking model of scispaCy *en_core_sci_lg* package (Neumann et al., 2019). The scispaCy NER model was trained from nine datasets covering a diversity of entities in several biomedical domains. The scispaCy Entity Linking model is capable of linking an entity to UMLS concepts and concept *aliases* which include common names of drugs and alternative spellings.

2.3 Ensemble Learning

Lastly, to predict whether each tweet contained a medication or dietary supplement mention, we used an ensemble of BERT-based linear classifiers. BERT (Devlin et al., 2018) is a pre-trained contextual word representation encoder based on transformers and can be fine-tuned using task-specific layers in addition to BERT for a specific task. BERT has been used in state-of-the-art systems for several Natural Language Processing (NLP) tasks, including winning systems of the SMM4H 2019 Shared Task (Weissenbacher

Model	Bio+Clinical BERT			BERT-Base			BERT-Large		
Method	Single	10-fold ensemble		Single	10-fold ensemble		Single	10-fold ensemble	
		Average	Hard Voting		Average	Hard Voting		Average	Hard Voting
F1	83.87	83.58	73.68	81.16	84.38	79.45	73.17	***84.85***	79.37
Precision	***96.30***	87.5	68.29	82.35	***93.10***	76.32	63.83	90.32	89.29
Recall	74.29	80.00	80.00	80.00	77.14	82.86	85.71	80.00	71.43

Table 1: Performance on SMM4H 2020 validation dataset of the single models and 10-fold ensemble models using *individual* BERT models

Model	Bio+Clinical BERT and BERT-Base	Bio+Clinical BERT and BERT-Large	BERT-Base and BERT-Large	Bio+Clinical BERT, BERT-Large and BERT-Base
F1	84.85	***86.15***	85.71	83.08
Precision	90.32	93.33	***96.43***	90.00
Recall	***80.00***	***80.00***	77.14	77.14

Table 2: Performance on SMM4H 2020 validation dataset of multiple *combined-BERT* models in addition to 10-fold ensemble

et al., 2019b) which focused on using machine learning techniques to detect tweets containing adverse drug reactions (Task 1) or personal health mentions (Task 4).

Since the tweets in this shared task belong to the consumer health domain, they contain language characteristic of the general English, clinical, and biomedical domains. We experimented with three different pre-trained BERT models: two from the general English domain, *BERT-Base Uncased* and *BERT-Large Uncased* (Devlin et al., 2018), and one from the biomedical and clinical domains, *Bio+Clinical BERT* (Alsentzer et al., 2019). *Bio+Clinical BERT* uses *Bio-BERT* (Lee et al., 2020) which is trained on biomedical text, to initialize embeddings which are then trained on clinical texts. It was shown to outperform *Bio-BERT* and *BERT-Base Uncased* on three of five common Clinical NLP tasks (Alsentzer et al., 2019). We fine-tuned each of these BERT models using a linear classification layer with AdamW optimizer for four epochs with a batch size of 16 and a learning rate of 0.00001 in a 10-fold cross-validation setup before starting the ensemble process.

For each pre-trained word embedding model, *BERT-Base Uncased*, *BERT-Large Uncased* and *Bio+Clinical BERT*, we trained both a single model and 10-fold-ensemble model. In the single model, the training dataset was split into training and validating subsets with ratio 90:10, and the trained model was used to predict labels for tweets in the validation dataset. In the 10-fold-ensemble model, 10 different models were created from the data splits of 10-fold cross-validation. Therefore each separate model had its own training and validating subsets. We experimented with two ensemble methods to combine the outputs of these systems: (1) Average, in which we take the mean of predicted probabilities of each individual classifier and use argmax to obtain the class label, (2) Hard Voting, in which we select the majority class of the class labels predicted by each individual classifier.

To balance out the individual weaknesses of the three pre-trained BERT models, we also performed four more experiments which combined *Bio+Clinical BERT*, *BERT-Base Uncased*, and *BERT-Large Uncased* ensemble models into a single ensemble. We experimented by combining each pair, and all three models using the average ensemble method to combine their predictions.

2.4 Additional Training Data

After training our initial model, we performed an error analysis on the validation set to find reasons for false negative predictions. We found that although each false negative sample included a word or a phrase related to medications, their predicted positive probability was low. These samples typically contained either (1) medication-related words or phrases that were either unseen or rarely seen in the training dataset, or (2) contained medication-related words or phrases which were inconsistently labeled within the training dataset. To correct for these errors, we added data from the SMM4H Shared Task

Model	Bio+Clinical BERT	BERT-Base	BERT-Large	Bio-Clinical-BERT, BERT-Base	Bio-Clinical-BERT, BERT-Large	BERT-Base, BERT-Large	Bio-Clinical-BERT, BERT-Large, BERT-Base
F1	84.06	84.93	83.33	**89.55**	**89.55**	84.51	86.96
Precision	85.29	81.58	81.08	93.75	93.75	83.33	88.24
Recall	82.86	88.57	85.71	85.71	85.71	85.71	85.71

Table 3: Performance on SMM4H 2020 validation dataset of *individual* and *combined-BERT* models in addition to 10-fold ensemble using additional training data

2018 on classifying tweets mentioning drug names or dietary supplement (Weissenbacher et al., 2018) to the training set. This dataset contained labeled samples for most of the false-negatives predictions our system made and contained 4,975 positive and 4,647 negative tweets.

3 Results

3.1 System Development

Table 1 compares performance of the single models and 10-fold ensemble models on the SMM4H 2020 validation dataset using each pre-trained BERT model. For all 10-fold ensemble models, the F1 score of the average method was higher than that of the hard voting method, and the average method was either very similar to or higher than the single, non-ensemble model. Table 1 also shows that *BERT-Large Uncased* 10-fold-ensemble model achieved the highest F1 score at 84.85%, indicating it is the most efficient model using single BERT model. However, we found that combining multiple BERT models increased performance. Table 2 shows that when combining 10-fold ensemble models, the combination of *Bio+Clinical BERT* and *BERT-Large Uncased* performs the best with F1 score at 86.15%.

Tables 1 and 2 show performance when using only the SMM4H 2020 training data. Table 3 shows the performance of multiple average ensemble models using both the SMM4H 2020 and 2018 data for training. Both *Bio+Clinical BERT* and *BERT-Large Uncased* ensemble model and *Bio+Clinical BERT* and *BERT-Base Uncased* ensemble model achieved the highest F1 score at 89.55%. This performance increased by 3.4% compared to the best *Bio+Clinical BERT* and *BERT-Large Uncased* ensemble model without additional data in Table 2. This shows that including related training data improved performance.

3.2 System Submissions

We made three system submissions for test dataset evaluation. Each of these systems used the best configuration indicated by our validation set experiments, and consisted of an ensemble of 20 *BERT-Large Uncased* and *Bio+Clinical BERT* classifiers using the average ensemble method. Each system was trained on a combination of the training and validation datasets from SMM4H 2020, and the training dataset of SMM4H 2018. The submissions differed in their use of the pre-filtering phase. The first submission used pre-filtering for training and evaluation. The second submission did not use pre-filtering for training, but used it for evaluation. The third submission did not use pre-filtering at all. The third submission obtained the highest F1 score on the test dataset of Shared Task 2020 at 85.35%, followed by the second variant and the first variant with F1 score at 85% and 83% respectively. These results indicate that pre-filtering does not help the system classify tweets on the test dataset. The third submission is the highest performing system in Task 1 of the SMM4H Shared Task 2020. In addition, its F1 score outperformed that of the baseline system `Kusuri` by 6.56%.

4 Conclusion

In this work, we developed an ensemble tweet classifier based on BERT combined with pre-processing and pre-filtering steps. The system was evaluated on a natural, highly imbalanced dataset, achieved the highest F1 score in Task 1 of the SMM4H Shared Task 2020. We found that the best system configuration consisted of an ensemble of 20 BERT-based classifiers built using *BERT-Large Uncased* and *Bio+Clinical BERT* ensembled using an average method. We also found that including related training data improved performance, and that the pre-filtering step did not improve performance. For future work, we plan to experiment with different neural network classification layers in addition to BERT outputs.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*.

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*.

Ari Klein, Ilseyar Alimova, Ivan Flores, Arjun Magge, Zulfat Miftahutdinov, Anne-Lyse Minard, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Davy Weissenbacher, 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*.

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.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy, August. Association for Computational Linguistics.

Davy Weissenbacher, Abeed Sarker, Michael J. Paul, and Graciela Gonzalez-Hernandez. 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, Brussels, Belgium, October. Association for Computational Linguistics.

Davy Weissenbacher, Abeed Sarker, Ari Klein, Karen O'Connor, 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 O'Connor, Michael J. Paul, and Graciela Gonzalez-Hernandez. 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, Florence, Italy, August. Association for Computational Linguistics.