Neural Network to identify personal health experience mention in tweets using BioBERT embeddings

Shubham Gondane*
Arizona State University
sgondane@asu.edu

Abstract

This paper describes the system developed by team ASU-NLP for the Social Media Mining for Health Applications(SMM4H) shared task 4. We extract feature embeddings from the BioBERT (Lee et al., 2019) model which has been fine-tuned on the training dataset and use that as inputs to a dense fully connected neural network. We achieve above average scores among the participant systems with the overall F1-score, accuracy, precision, recall as 0.8036, 0.8456, 0.9783, 0.6818 respectively.

1 Introduction

There has been an increase in the use of social media worldwide in recent years, which provides an abundance of data available and an exciting opportunity to build and improve biomedical and public health applications. The Social Media Mining for Health Applications (SMM4H) Workshop 2019 (Weissenbacher et al., 2019) proposed four tasks. We have focused on task 4, which was the most interesting. The task is to classify whether the tweet contains personal health mention as opposed to a general discussion of the topic. The training data consisted of tweets related to the flu. The system is evaluated on tweets related to flu and a second health domain across two contexts.

1.1 Data Description

The organizers provided two datasets across different contexts, but both in the flu domain. The first dataset had 1046 records of flu infection, but around 1023 tweets were available for download. The flu vaccination dataset had around 9800 records out of which only 6659 were available for download. The combined dataset had 7682 tweets in total.

1.2 Related Work

Much previous work has focused on tracking and monitoring diseases on social media. Identifying various health ailments in social media by (Paul and Dredze, 2011) introduced a topic model based system using LDA to discover health mentions. Previous work done on creating generalizable classifiers have used traditional machine learning based approaches. (Yin et al., 2015) have developed a scalable system by training classifiers on a dataset of 34 health topics. They created a general health classifier using standard SVM with an accuracy of 77 percent. More recently, (Karisani and Agichtein, 2018) developed a system called as WESPAD that combines lexical, syntactic, word embedding-based, and context-based features. The authors report that the system can generalize from a few examples by automatically distorting the word embedding space to detect the accurate health mentions most effectively.

1.3 Preprocessing

The challenge in this task is to train a model on one disease domain and test on another, so it is important to make sure the model does not learn disease-specific characteristics. One way to ensure this is to mask specific terms like flu or influenza mentions with an AILMENT tag. A list of all flu-related terms was created using a pre-trained Word2Vec model for Twitter (Godin et al., 2015) to find similar terms to flu. The list was expanded using human knowledge and ConceptNet[1] (Speer et al., 2017). This list of terms was used to mask all the flu mentions in the dataset.

Additionally we use the preprocessing library Ekphrasis to clean the tweets. (Baziotis et al., 2017).

* The author is advised by Dr. Chitta Baral at Arizona State University.

[1] www.conceptnet.io

Proceedings of the 4th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 110–113
Florence, Italy, August 2, 2019.

- All @user mentions were replaced by @user tag.
- All HTTP URLs were replaced by URL tag.
- Hashtags were preprocessed by removing the # symbol and keeping the words.
- Emojis, dates, numbers, etc. are removed.

2 Experiments

Language models like BERT (Devlin et al., 2018) and OpenAI GPT-2 (Radford et al., 2019) have achieved state of the art performances in various NLP tasks. Such models that are trained on large datasets can be fine-tuned on smaller datasets to achieve good scores on various NLP tasks. BioBERT (Lee et al., 2019), a domain-specific language representation model designed for biomedical text, is built using BERT architecture. Our system is built using transfer learning approach by fine tuning on the given dataset using the BioBERT model.

2.1 Fine-tuning

The fine-tuning process involves creating a train and dev set in the format provided by the data processor in the BERT/ BioBERT model. The BERT-base uncased model is used for the experiments [2]. The model is then trained on a sentence classification task end to end using the default parameter values provided by the authors. Fine-tuning on smaller dataset results in a high variance in the dev set accuracy. So the model with the best result on the dev set is selected after five iterations of the fine-tuning process. This process is applied for fine-tuning both the BERT and BioBERT v1.0 models. BioBERT produced a slightly better model with the difference in dev set accuracies of the final BERT and BioBERT fine-tuned models was less than 2 percent.

We also experimented with fine-tuning without doing any preprocessing on the tweets. As expected, the performance decreased quite significantly because BERT does token level masking and presence of URLs, hashtags, and @user-mentions makes this token level prediction more difficult.

[2]The BioBERT model v1.0 used in this system is also based on the BERT-base model.

2.2 Dense Neural Network Model

The BERT model can also be used for extracting features by fine-tuning the model and extracting the fixed contextual representations of each token. These features can be used in conjunction with other features in a different model. Fine-tuning is essential because the training set for these models is quite different from the dataset for this task. It helps to adjust the model weights that are closer to the target domain.

The BERT/BioBERT model adds two tokens in each input line - a CLS token in the beginning and SEP token at the end. Two feature embeddings are extracted in the following manner. In one case, we mask the flu-mentions, and in the other, the flu-mentions are kept as it is. The embedding for the CLS token is extracted by concatenating the weights of the last four layers of the BioBERT model. In their paper (Devlin et al., 2018), the authors state that concatenating last four layers gives the best result.

These embeddings are used as the input layer to a dense neural network with two hidden layers. We tried using these embeddings separately and also concatenated the two. The concatenated embedding performed slightly better than just using either of them separately. The final network has a 6144-dimensional input layer followed by two hidden layers of 512 and 128 dimensions, respectively. A dropout layer is added between the two hidden layers, and the hyperparameters are tuned accordingly.

3 Results and Discussion

Since the test set contained tweets related to undisclosed context we created a list of health concerns discussed on Twitter from previous research work (Daughton et al., 2018) (Paul and Dredze, 2014) (Dalrymple et al., 2016) (Khatua et al., 2019) done on exploring health-related tweets for analysis. This extensive list was used to mask the tweets of test set so that the masked embeddings make some contribution to the classification.

The system we used for this task shows that language models like BERT and BioBERT can be fine-tuned on a small dataset of tweets and still achieve promising results on test set where the health concern was similar to the training set. Transfer learning across different domains is still a challenging task as it is evident from the results.

It is interesting given that these models are

Model	Acc	F1	P	R
BERT fine-tuned without preprocessing	0.82	0.8	0.79	0.82
BERT fine-tuned	0.86	0.85	0.86	085
BioBERT fine-tuned	0.87	0.85	0.87	0.85
BioBERT unmasked embeddings	0.90	0.89	0.94	0.86
BioBERT masked embeddings	0.91	0.91	0.97	0.85
BioBERT masked and unmasked embeddings	**0.93**	**0.92**	**0.97**	**0.88**

Table 1: Accuracy, F1 score, Precision and Recall results on training data using different models and embeddings.

Test set	Acc	F1	P	R
health concern overall	0.84	0.80	0.97	0.68
health concern condition 1	0.92	0.92	0.98	0.86
health concern condition 2	0.69	0.51	0.91	0.35
health concern condition 3	0.80	0.59	1	0.42

Table 2: Final Accuracy, F1 score, Precision and Recall scores on the test set for the best performing run submitted.

trained on Wikipedia or biomedical text that how well they perform on tweets as tweets often contain misspellings, sarcasm, and slangs. It would also be interesting to see if the model can perform better if we had a BERT model trained on tweets or if we had a larger training dataset. This model could possibly be further improved by using additional data and the use of other textual and semantic features combined with the embeddings from the BioBERT model or trying different architectures.

References

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada. Association for Computational Linguistics.

Kajsa E Dalrymple, Rachel Young, and Melissa Tully. 2016. facts, not fear negotiating uncertainty on social media during the 2014 ebola crisis. *Science Communication*, 38(4):442–467.

Ashlynn R Daughton, Michael J Paul, and Rumi Chunara. 2018. What do people tweet when theyre sick? a preliminary comparison of symptom reports and twitter timelines. In *ICWSM Social Media and Health Workshop*.

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*.

Fréderic Godin, Baptist Vandersmissen, Wesley De Neve, and Rik Van de Walle. 2015. Multimedia lab @ ACL WNUT NER shared task: Named entity recognition for twitter microposts using distributed word representations. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 146–153, Beijing, China. Association for Computational Linguistics.

Payam Karisani and Eugene Agichtein. 2018. Did you really just have a heart attack?: towards robust detection of personal health mentions in social media. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 137–146. International World Wide Web Conferences Steering Committee.

Aparup Khatua, Apalak Khatua, and Erik Cambria. 2019. A tale of two epidemics: Contextual word2vec for classifying twitter streams during outbreaks. *Information Processing & Management*, 56(1):247–257.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Michael J. Paul and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. In *ICWSM*.

Michael J Paul and Mark Dredze. 2014. Discovering health topics in social media using topic models. *PloS one*, 9(8):e103408.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1:8.

Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael Paul, and Graciela Gonzalez-Hernandez. 2019. Overview of the fourth Social Media Mining for Health (SMM4H) shared task at ACL 2019. In *Proceedings of the 2019 ACL Workshop SMM4H: The 4th Social Media Mining for Health Applications Workshop & Shared Task*.

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.