Implementing BERT to Improve Personal Health Mention Detection

Claira Kauffmann, Akhila Ganti, Rachael Phillips
School of Information, UC Berkeley

Abstract

Personal health mention detection, predicting whether a text is a report of a health condition, has become even more prevalent as "social media has become a substitute for social interaction. This has caused an increase of medical and clinical-related information on the web" (Khan). A unique use case occurred during COVID-19, when all interactions were brought online, many organizations tried to track the spread of the virus from social media, specifically X, formerly Twitter. Similarly, we aim to improve the personal health mentions predictions using more current models and pre-trained datasets. Implementing these newer models, such as the BioClinical BERT, we are seeing an improvement of 49.46%. We will show that utilizing more current BERT models to process the Twitter data, our PHM detection will improve.

1 Introduction

Our goal is to improve results of the model and methodology developed in the 2019 research paper, *Figurative Usage Detection of Symptom Words to Improve Personal Health Mention Detection*. We are using this paper as inspiration and aim to improve the accuracy of Personal Health Mention Detection. Personal health mention detection aims to detect whether or not a text contains a personal health mention, which is the report of a health condition and is abbreviated as PHM. A PHM is often detected by identifying 'symptom words' in the text, such as *'sneezes'* or *'coughing'*. Personal Health Mention Detection (PHMD) in social media has been around for years, however, its relevance will only increase as social media, and communication through social media, becomes more ingrained in our everyday lives and the amount of data generated from these interactions increases.

Previous PHM detection models would often have errors due to figurative language. As a result, the 2019 research paper aimed to improve previous PHMD models by first identifying if the text had Figurative language (Iyer 2019). Figurative Language does not use a word's strict or realistic meaning. An example of figurative language that also includes 'symptom words' is *'When Paris sneezes, Europe catches cold*'. This is figurative language and it is not a PHM. It could most likely be misclassified as it contains symptom words. Although it sounds straightforward, both tasks, accurately identifying figurative language and identifying personal health mentions, are inherently difficult. Both tasks, when used in a CNN model, improved Personal Health Mention detection by 2.2%.

Convolutional Neural Networks (CNNs), while showing some effectiveness in identifying figurative speech, face limitations, particularly when dealing with complex textual scenarios in social media. Despite incorporating GloVe embeddings trained on tweets, CNNs still struggle to grasp the complex syntactical nuances and long-range dependencies of natural language.

BERT would be a better model to consider for this multi classification problem as it allows for

contextual understanding of the information. BERT architecture's attention mechanism which allows focus on relevant parts of the input text will improve score and accuracy. Additionally, pre-trained models can be used and fine tuned on specific tasks.

2 Background

Tracking Personal Health Mentions on Social Media can help public health departments, such as the World Health Organization, monitor trends and track the spread of epidemics in the population. For example, a similar paper and dataset was used to track the spread of COVID-19 and could be used as a crucial method for tracking the pandemic conditions in real time (Luo 2022).

In the 2019 study, they found that by adding Figurative Language Detection to their model, their F-1 score increased by an average of 2.21%. We are using this study as inspiration and aim to improve the overall F-1 score, not only through adjusting Figurative Usage Detection.

In addition to the Figurative Language Usage detection used in the 2019 research paper, there has been research into improving PHM detection. One such research study tried to improve results by using permutation based word representation learning. They tried to capture the context of disease words efficiently to improve the performance of the classifier (Khan). Ultimately, they achieved a 5.5% improvement in their F-score compared to the public benchmark dataset.

Our goal is to improve the F-1 score against our baseline model that we will discuss further later on in this paper.

3 Relevant Dataset

We will be using a data set from X, formerly Twitter, of 7,192 English Tweets that have been labeled. This dataset is often referred to as 'PHM2017'. It consists of ~ 5837 Tweets across six diseases and conditions. As a part of the experiment, Tweets were manually annotated with labels (self-mention, other-mention, awareness, non-health). As noted in the 2019 research paper, there was an imbalance in the class labels.

The dataset has been updated since 2017 and we are using a combination of Tweets from 2017 and 2019 dataset. The researchers from the 2019 paper assigned the samples to one of three classes, shown in Table 1, which will be our output of interest.

The aforementioned dataset contained only the Tweet ID numbers and required X's, formally Twitter's API, however, because of the API's limitation, we were only able to pull 9,733 tweets. This is more than the 2017 dataset, but fewer Tweets than the 2019 dataset.

In this data, we have a balance between classes, though we note a class imbalance in the subset of the health terms. We will be using this dataset for all of our models.

ID	Type	Example
0	Figurative Mentions	Chuck Norris gives the sun cancer.
1	Other Mentions	Nationalism is a cancer.
2	Health Mentions	I'm a single mom to 4 kids and this is my third time facing cancer. I have Ewings Sarcoma, and there is no specialist anywhere near us.

Table 1: Example of Tweets and classifications

4 Methods

4.1 Baseline CNN

In our initial exploration, we aim to replicate a CNN model from the 2019 study, focused on 'Figurative Usage Detection of Symptom Words'. Following the specified hyperparameters in the research, we built a CNN model to establish a baseline to improve upon classifying medical Tweets into three categories: figurative, personal health mentions, and general health references.

During this evaluation process, using 10-fold Cross Validation, we noticed certain limitations in the CNN model. A significant shortcoming was the frequent misclassification of Class 2 tweets as Class 0, likely due to CNNs narrow viewing window.

The shortcomings from CNN's take on a greater weight in the realm of Personal Health Mention Detection (PHMD). For example, in a Tweet such as "*I can't believe the team traded Smith; it's like they're committing suicide*," the word "*suicide*" is used metaphorically to describe a baseball team's decision, not as a personal health crisis. In the aforementioned example, the key context – a reference to a baseball team – lies outside the CNN's local window of focus, leading to potential misclassification.

The subtlety between literal and figurative language, often reliant on the broader narrative rather than immediate word surroundings, and often some world knowledge or understanding of current events, poses a significant challenge for CNNs, and suggests the need for more sophisticated options. However, this gave us a suitable starting point as we sought enhancements that could potentially refine the model's accuracy.

CNN Baseline Architecture:

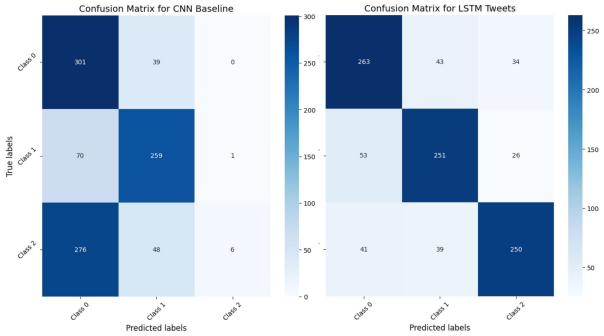
Our Convolutional Neural Network (CNN) was built using TensorFlow and integrated GloVe embeddings for word representation. The GloVe embeddings used were sourced from glove.twitter.27B.100d.txt, providing a 100-dimensional vector for each token. The architecture comprised three convolutional layers, each followed by max-pooling and dropout layers.

Training and Validation

The model was trained for 32 epochs with a batch size of 128, using 10-fold cross-validation to ensure robustness and generalizability.

4.2 LSTM and spaCy

To address the limitations observed in the CNN model, we integrated a Long Short-Term Memory (LSTM) network. Our hypothesis was that LSTM, especially when combined with GloVe embeddings, would better capture the temporal and contextual subtleties of language found in Tweets because they have a longer memory.



We did struggle with the LSTM model overfitting. We hypothesized that if we could help the model focus on syntactic structure rather than just the content, we might get it to generalize better, especially given the class imbalance in our dataset among the different health terms. First, we tried masking out the health term used in the tweet and then trained our LSTM model on the modified text. The F1 score only dropped a few percentage points and made us confident that the imbalance in health

terms was not hurting our model's generalizability.

We believed the LSTM was picking up on broader syntactical differences in the figurative Tweets. This led us to test spaCy, a Python library for natural language processing. With this library, we extracted part-of-speech tags from the tweets. For instance, a tweet like "*Gun violence is a cancer on our society*" would be transformed into a sequence of part-of-speech tags, such as "[NOUN] [NOUN] [VERB] [DET] [NOUN] [ADP] [DET] [NOUN]". We plan to sample the results here to determine if our inference is correct.

4.3 Utilizing BERT

Next, we plan to utilize BERT models to improve our results and aim to improve the models accuracy as a whole, not through only Figurative Usage Detection done as the 2019 research paper.

In general, BERT architecture has a bidirectional context understanding, a self-attention mechanism and is pre-trained on large text corpora which collectively contribute to BERT's effectiveness in capturing intricate language patterns, leading to its superiority in various NLP tasks compared to the earlier models.

Our first BERT model will be a basic BERT model with no fine-tuning. It will be used as a base to compare our more advanced BERT model. We have selected two Bio Clinical and Clinical pre-trained models from the Hugging Face library. We will be building a simple classification model and use the pooler output for multi classification.

As BERT models are pre trained on data, we expect our base BERT model to perform better than a CNN on text data taken from Tweets.

4.4 Pretrained BERT

We will compare these results to a BioClinical BERT model that has been pre-trained on more medical-related text. Specifically, the Bio_ClinicalBERT model, from HuggingFace, was trained on all notes from MIMIC III, a database containing electronic health records from ICU patients at the Beth Israel Hospital in Boston, MA (Alsentzer). We expect to see improvement in the prediction compared to the baseline model as more symptoms words would be accurately identified.

Last, we will compare these to a BERT using CORe(Clinical Outcome Representations), which was introduced for clinical outcome predictions from admission notes using self-supervised knowledge integration. It incorporates the ICD (International Classification of Disease codes) and clinical notes text data (Aken 2021)

To compare the performance of these models, we will be reporting F-1, precision and accuracy scores. F-1 scores were also commonly used to communicate model improvements in previous Personal Health Mention Detection research and we will remain consistent.

5 Results and Discussion

The results of each model can be found in Table 2. As predicted, our F scores improved through each iteration of models.

Model	F-1	Precision	Accuracy
CNN	0.58	0.59	0.64
CNN + LSTM	0.78	0.79	0.78
BioClinical BERT	0.87	0.87	**0.8706**

Table 2: Performance of PHM Detection

5.1 Baseline CNN

After testing our CNN model with similar parameters in 2023, we immediately saw improvements in our evaluation metrics which

can be explained by our use of GloVe 200. GloVe can generally be trained on smaller datasets as opposed to Word2vec which requires a large amount of training data (Success Team). Additionally, GloVe 200 has been trained on Tweets. As models tend to improve throughout the years, the training datasets do as well. The 2019 research papers average was 49.26 and the result of our CNN was 56.

Performance

The model achieved an F1 score of 58.21, with the following class-specific performance:

Class 0 (Figurative Health Mention): TP: 301, FP: 346, FN: 39
Class 1 (Personal Health Mention): TP: 259, FP: 87, FN: 71
Class 2 (General Health Mention): TP: 6, FP: 1, FN: 324

When sampling the results of our CNN model, we noticed Class 0 (Figurative Mentions) has many instances where it is falsely predicted as a Figurative Mention when it is actually a 'Other Mention (96 Tweets) or a 'Health Mention' (91 Tweets).

5.2 LSTM

Training the LSTM solely on these tag sequences, we were surprised to find the model still performed effectively, confirming its ability to discern structural patterns apart from contextual meaning. As expected, the LSTM model with GloVe embeddings showed a marked improvement in performance, achieving an F1 score of 76.44, significantly surpassing the CNN's F1 baseline by 18.23 percent.

5.3 CNN + LSTM + spaCy

We experimented with a combined CNN and LSTM model to boost the F1 score by allowing them each to specialize on their strengths. The CNN part of the model was trained on the original Tweet texts, while the LSTM focused on the part-of-speech patterns pulled using spaCy, and then their vectors were combined.
This hybrid model yielded our highest F1 score yet among non-transformer models, reaching 78.2. However, the small increase in performance compared to the added model complexity did not seem to be a good trade-off.

Performance

The class-specific performance was as follows:

Class 0 (Figurative Health Mention): TP: 263, FP: 94, FN: 77
Class 1 (Personal Health Mention): TP: 251, FP: 82, FN: 79
Class 2 (General Health Mention): TP: 250, FP: 60, FN: 80

5.4 BERT models

These experiments, particularly the surprising effectiveness of LSTM on part-of-speech sequences, guided us towards exploring transformer models, specifically BERT. Our rationale was that BERT's advanced capabilities in understanding context and language structures would potentially offer even more significant improvements in Tweet classification.

BERT Model	F-1	Accuracy
Base BERT	0.61	0.6143
COREBERT	0.86	0.8597
BioClinical BERT	**0.87**	**0.8706**

Table 3: Performance of PHM Detection - BERT models

Baseline BERT

Our baseline BERT model has a F-1 score of 0.61, which was an improvement of 3.1% from the baseline. Without any fine-tuning, it appears our BERT model was understanding the Tweets which resulted in a better accuracy and F-1 score.

BioClinical BERT

Improving on the base BERT model, when implementing a BioClinical BERT model, our F-1 score increased to 0.87, which was an improvement of 49.46% from the baseline and 45% from the base BERT model. The use of the BioClinical pretraining made a difference. We continued to experiment with the baseline model through various adjustments to the hyperparameters. Additionally, we compared these results with another Clinical BERT model which resulted in similar performance. Next, we continued with the Bio Clinical pre-trained model by unfreezing layers, adjusting learning rates, batch rates and number of epochs for training the model. A summary of the fine-tuning results can be found in Table 4.

We learned the effects of changing the hyper parameters and how it improves or degrades the performance of the model. The frozen layers are responsible for adapting the model's representations to better suit the nuance of the task. Reducing the number of epochs from 5 to 3 improved the scores as it prevented overfitting during training while increasing the learning rate from 0.00005 to 0.0005 caused difficulty in converging to the optimal solution. Increasing batch size from 8 to 20 showed remarkable improvement in the F1 score.

Effects of Fine-Tuning BERT Models	**F1**
BaseBERT compared to BioClinical BERT with frozen layers	0.60
BioClinical BERT compared to Clinical BERT with unfrozen layers	0.86
BioClinical BERT - Reducing the number of epochs from 5 to 3 to reduce overfitting	0.87
Bio Clinical BERT - Changing learning_rate from 0.00005 to 0.0005	0.35
Bio Clinical BERT - Changing batch size from 8 to 20	0.86

Table 4: Summary of Fine-tuning on BERT

6 Sampling Results

Without solely focusing on improving Figurative Usage Detection, our BERT model was able to accurately detect Figurative language from the Tweet, *"My dog got out today and I seriously had a miny heart attack and had the biggest knot in my stomach."*

Regarding errors with the BERT prediction, we are seeing consistent errors with prior research where if Health Words are included in the text, they are tagged incorrectly, "*Generic Health - Stages of Alzheimer's Disease and What to Expect https://t.co/d0ECr0EEGj and Personal Health - @teddavid @Mediaite He had Alzheimer's, which is formally recognized as a mental illness. https://t.co/RLzXSCmVOc"*

Through these various experiments and fine tuning, we learned aspects of model experimentation and tuning with the BERT models. Additionally, we learned how cloud resources for memory, GPU, etc are a big consideration for the model performance during the training and validation phases.

7 Conclusion

In this paper, we utilized BERT models, trained on BioClinical data, to improve Personal Health Mention Detection. Compared to our baseline, we were able to improve the F-1 score by 49.46% . Our main contribution is that we have developed and tested an approach to personal health mention detection that is more straightforward while maintaining a high accuracy.

References

Iyer, Adith, et al. “Figurative usage detection of symptom words to improve personal health mention detection.” *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, July 2019, https://doi.org/10.18653/v1/p19-1108.

Luo, Linkai, et al. “Covid-19 personal health mention detection from tweets using dual convolutional neural network.” *Expert Systems with Applications, vol. 200,* 2 Apr. 2022, p. 117139, https://doi.org/10.1016/j.eswa.2022.117139.

Alsentzer, Emily. “Bio_ClinicalBERT · Hugging Face.” *Emilyalsentzer/Bio_ClinicalBERT · Hugging Face*, huggingface.co/emilyalsentzer/Bio_ClinicalBERT. Accessed 27 Nov. 2023.

Khan, P.I., Razzak, I., Dengel, A., Ahmed, S. (2020). Improving Personal Health Mention Detection on Twitter Using Permutation Based Word Representation Learning. In: Yang, H., Pasupa, K., Leung, A.CS., Kwok, J.T., Chan, J.H., King, I. (eds) Neural Information Processing. ICONIP 2020. Lecture Notes in Computer Science(), vol 12532. Springer, Cham. https://doi.org/10.1007/978-3-030-63830-6_65

Aken, Betty Van, and Jens-Michalis Papaioannou. “DATEXIS/Core-Clinical-Diagnosis-Prediction · Hugging Face.” DATEXIS/CORe-Clinical-Diagnosis-Prediction · Hugging Face, Association for Computational Linguistics, 2021, huggingface.co/DATEXIS/CORe-clinical-diagnosis-prediction.

Success Team. “Word2vec vs Glove.” Speak Ai, 13 Feb. 2023, speakai.co/word2vec-vs-glove/#:~:text=Word2vec%20requires%20a%20large%20amount,better%20suited%20for%20larger%20applications.

Appendices

Appendix A Additional CNN Specifications

CNN Baseline Architecture:
Our Convolutional Neural Network (CNN) was built using TensorFlow and integrated GloVe embeddings for word representation. The GloVe embeddings used were sourced from glove.twitter.27B.100d.txt, providing a 100-dimensional vector for each token. The architecture comprised three convolutional layers, each followed by max-pooling and dropout layers:

1 . First Layer: Conv1D with 100 filters, kernel size of 3, and ReLU activation, followed by a MaxPooling1D with pool size 2 and a Dropout layer with a dropout rate of 0.2.
2. Second Layer: Conv1D with 4 filters, kernel size of 4, and ReLU activation, followed by MaxPooling1D with pool size 2 and a Dropout layer with a dropout rate of 0.3.
3. Third Layer: Conv1D with 4 filters, kernel size of 5, and ReLU activation, followed by MaxPooling1D with pool size 2 and a Dropout layer with a dropout rate of 0.5.