The Effect of Using Sentiment Analysis Features on Mental Health Classification

Fagun Mehulbhai Raithatha, Daniel Stapleton, Michal Lange and Thushara Manjari Naduvilakandy
Department of Computer Information Science, IUPUI, Indianapolis
{fraitha, dpstaple, milange ,tnaduvil}@iu.edu
Prof. Hyeju Jang
Department of Computer Information Science
IUPUI, Indianapolis
hyejuj@iupui.edu

Abstract

The accurate identification of patients in need of mental health treatment is a pressing concern in healthcare. In this study, we aim to investigate the impact of incorporating sentiment analysis into a BERT-based classification model to improve the accuracy and effectiveness of mental health diagnosis. Our hypothesis is that integrating emotional analysis into the feature vector of the model will provide valuable insights into patients' mental health needs, thereby enhancing the classification process's reliability. We plan to compare the performance of a classification model with and without sentiment analysis on varying sizes of training sets to determine the optimal approach. Furthermore, we intend to investigate the differences in results when using different sentiment analysis models trained on Twitter data or Facebook posts. Our study aims to provide a more comprehensive understanding of the potential benefits of integrating sentiment analysis into mental health diagnosis and to contribute to the development of more accurate and effective diagnostic tools.

Keywords: BERT, KW-ATTN

1 Introduction

Our objective is to create a classification model that can accurately determine if a patient requires mental health treatment. Our idea is to get feature input from the BERT model and also the Sentiment-BERT model. Our hypothesis is getting sentiment behind inputs can be beneficial in classifying if a person is disturbed or not.

Our study aims to investigate the impact of incorporating this additional feature on the performance of the classification model. By combining the sentiment analysis and BERT models, we hope to improve the accuracy and effectiveness of the classification process, thereby providing a more reliable tool for identifying patients in need of mental health treatment.

To elaborate further, sentiment analysis involves analyzing and classifying emotions in text data. By integrating this information into our model, we can better understand the emotional state of the patient, which can provide valuable insights into their mental health needs. Additionally, the use of BERT, a state-of-the-art language processing model, enables us to incorporate a range of textual features that are crucial for accurately predicting mental health needs. By leveraging both these models, we aim to create a classification model that is highly effective in identifying patients who require mental health treatment.

Our objective is to assess and contrast the effectiveness of two distinct methods for classification modeling: a simple classification model without sentiment analysis and a classification model that uses sentiment analysis. Our aim is to investigate how well each of these models performs on varying sizes of training sets. By doing so, we hope to determine which approach is more effective in different scenarios. We want to investigate the minimum amount of data required to achieve optimal results. We intend to compare the performance of these two approaches to gain a better understanding of the impact of sentiment analysis on classification modeling, and the potential benefits it can provide. Furthermore, we plan to incorporate different sentiment analysis models to observe the differences in results when using sentiment analysis models trained on Twitter data or those trained on facebook posts.

2 Related Work

The use of natural language processing (NLP) for sentiment analysis has been a growing area of research in recent years. Various models have been developed and fine-tuned for specific domains, such as social media. One such model is a fine-tuned Bert model for sentiment analysis on Twitter data. This model has been used to analyze sentiment in social media texts, but its potential for classifying mental health data has not been explored until recently. The current work aims to explore the effectiveness of this fine-tuned Bert model in conjunction with a Distilbert model for mental health classification tasks.

Transfer learning has been a popular approach in NLP, and many studies have shown that it can improve the performance of models for various tasks. This project draws inspiration from the work of (S and Antony, 2022) and (Rukhma Qasim and Almazroi, 2022), which have successfully applied transfer learning using a Bert model for sentiment analysis and text classification tasks. This project aims to apply similar transfer learning techniques to a mental health dataset to detect toxic language and flag it.

Ensemble methods have also been widely used in NLP for combining multiple models to improve performance. The project members looked at an ensemble method used in (Mayur Wankhade, 2022), which combines a Bert model with a bi-directional LSTM model for aspect-based sentiment analysis. The goal for us was to explore if a similar ensemble method could improve the classification of mental health data.

To combine the outputs of the fine-tuned Bert model and the Distilbert model, an attention layer was used. The project drew inspiration from the work of (Jang et al., 2021), which proposed a knowledge-infused attention mechanism to combine multiple sources of information for text classification. The attention layer was used to concatenate the outputs of the two models and examine the influence of the fine-tuned Bert model on the classification of the mental health data.

3 Dataset

We obtained a corpus related to mental health from Kaggle, which contains comments written by diverse individuals, and the comments have been categorized as either requiring medical attention due to their harmful nature, or as non-toxic.The corpus is a collection of 27972 labeled instances related to people with anxiety.The mental health dataset has an equal distribution of cases, with 51 percent of the instances labeled as 1, indicating that some of the verbiage used in the post warrants a second look or could be construed as a potential for self-harm or the harming of others, while the remaining 49 percent are labeled as 0 which means that there were no flags raised by the verbiage used in the post. For example:

- "nothing look forward life i dont many reasons keep going feel like nothing keeps going next day makes want hang myself" - Label:0
- "music recommendations im looking expand playlist usual genres alt pop minnesota hip hop steampunk various indie genres artists people like cavetown aliceband bug hunter penelope scott various rhymesayers willing explore new genres artists such anything generic rap the type exclusively sex drugs cool rapper is rap types pretty good pop popular couple years ago dunno technical genre name anyways anyone got music recommendations favorite artists songs" - Label:1

3.1 Pre-Processing

Once the dataset was analyzed, we proceeded to carry out preprocessing procedures on it such as:

- Convert text to lower case
- Tokenizing
- Stemming / Lemmatization
- Dealing with garbage/null data if any
- Dealing with special characters if any

4 Methodology

This project will utilize two sets of models. The first set features our baseline implementations, differentiated by two different versions of DistilBERT; the original and a fine-tuned version for sentiment analysis. The sentiment-analysis tuned DistilBERT was trained on a twitter dataset, which mirrors in style our dataset, hence the justification for its usage. In both base models the final CLS token output is passed through feed-forward layers for the final classification.

The target model includes both versions of DistilBERT, as well as the utilization of all output tokens. The combination of both tokens will use a previously developed method of word-concept layer attention. This is a two-step attention process that first creates a composition between corresponding pairs of tokens from both DistilBERT models, then the composed tokens will be merged into a single representation. The original paper on which this method based off of develops further representation by constructing a final vector from sentence level representations. However, due to the relatively short input texts of the dataset, this step will be omitted.

$$u_t : \tanh\left(W_\alpha \left[h_t, h_t^{sa}\right] + b_\alpha\right)$$

$$p_t : \sigma\left(w_p \left[h_t, h_t^{sa}\right] + b_p\right)$$

$$\alpha_t : \frac{\exp\left(u_t^T u_\alpha\right)}{\sum_t \exp\left(u_t^T u_\alpha\right)}$$

$$s : \sum_t \alpha_t \left(\left(1 - p_t\right) h_t + p_t h_t^c\right)$$

(Jang et al., 2021) $W_\alpha, b_\alpha, w_p, b_p, u_a$ are all learnable parameters of the model. $[h_t, h_t^{sa}]$ represents the concatenation of the equivalent token in both the base distilBERT and the sentiment-analysis tuned distilBERT. u_t is utilized for determining the significance of the token in the text relative to other tokens, denoted by a_t. p_t denotes the ratio between the two tokens for the composed representation.

Once the final vector s it will passed through a feed forward network for final classification.

Given enough time in the project we will also implement multitask learning in our model, incorporating a combined loss function for our task and the sentiment analysis task.

5 Results

5.1 Baseline Model

We possess a pair of baseline models: one trained on distill-uncased-bert, and the other model, which has been fine-tuned for sentiment analysis, trained on distilbert-base-uncased-emotion. Our objective is to evaluate the performance of each model and examine how they produce results when trained on varying sizes of data.

- Results of model trained on distill-uncased-bert:

Size of Training Data	F-1 score	Precision	Recall	Accuracy
10 percent	0.83	0.84	0.82	0.85
25 percent	0.89	0.92	0.87	0.89
50 percent	0.90	0.89	0.90	0.90
100 percent	0.89	0.86	0.92	0.89

- Results of model trained on distilbert-base-uncased-emotion:

Size of Training Data	F-1 score	Precision	Recall	Accuracy
10 percent	0.75	0.75	0.75	0.74
25 percent	0.82	0.78	0.87	0.80
50 percent	0.82	0.83	0.80	0.82
100 percent	0.85	0.80	0.90	0.84

5.2 Experiment Results

Upon evaluating our proposed model with the dataset, these are the outcomes that we have obtained.

Size of Training Data	F-1 score	Precision	Recall	Accuracy
10 percent	0.80	0.79	0.82	0.79
25 percent	0.88	0.84	0.92	0.88
50 percent	0.88	0.91	0.86	0.89
100 percent	0.91	0.91	0.90	0.91

5.3 Fine-Tune Experiment Results

- In order to optimize our model's performance, we made the decision to experiment with the cross-dimensions token parameter that is currently set to 768 in our model. We varied this parameter with different values and analyzed the resulting outcomes, which we are presenting here:

Cross-Dimension	F-1 Score
768	0.83
668	0.87
600	0.85
568	0.89
500	0.85
468	0.80

- We adjusted our model through a process of fine-tuning and experimentation with different parameter values, including cross-dimension values. Ultimately, we discovered that the parameter value of 568 produced the most optimal results. We then proceeded to test our model on various data sizes to compare its performance against our original model.

Training Data Size	F-1 Score
10 Percent	0.89
50 Percent	0.87
100 Percent	0.94

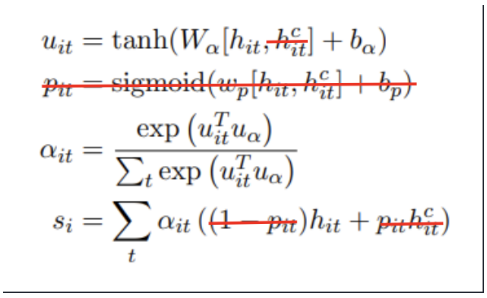
After conducting a new round of testing, we obtained the top-performing outcomes again, and the mean F-1 Score achieved was 0.923.

6 Discussion and Limitation

6.1 Comparing our model and a model with an attention layer

After achieving impressive results with our fine-tuned model, we wanted to ensure that these results were not simply due to the addition of an attention layer that analyzed individual token values from sentiment bert, but rather due to the model's overall functionality. Therefore, we created another model that also included an attention layer but did not take input from sentiment bert. We then made specific changes to this model and conducted further testing to compare its performance to our original fine-tuned model.

Changes we made to the model are as follows:



$$u_{it} = \tanh(W_\alpha[h_{it}; \cancel{h^c_{it}}] + b_\alpha)$$

$$\cancel{p_{it} = \text{sigmoid}(w_p[h_{it}; h^c_{it}] + b_p)}$$

$$\alpha_{it} = \frac{\exp\left(u_{it}^T u_\alpha\right)}{\sum_t \exp\left(u_{it}^T u_\alpha\right)}$$

$$s_i = \sum_t \alpha_{it}\left((\cancel{1 - p_{it}})h_{it} + \cancel{p_{it}h^c_{it}}\right)$$

Figure 1: Change in our initial model

It's important to note that the addition of an attention layer can greatly improve a model's performance, as it allows the model to focus on the most relevant parts of the input data. However, we wanted to make sure that the attention layer wasn't the sole reason for our initial success, but rather a complementary feature that enhanced the model's overall accuracy.

Parameters	Results
F-1 Score	0.91
Accuracy	0.91
Precision	0.87
Recall	0.95

It appears from our analysis that the outcomes obtained from above model without fine-tuning are quite comparable to those obtained from the model using sentiment BERT. Hence, our initial supposition that sentiment BERT would provide meaningful insights seems to be incorrect.

However, we cannot conclusively determine whether the similarity between the two sets of results was due to the introduction of an attention layer or the comprehensive evaluation that utilized every token for classification. Further investigation is required to ascertain the underlying cause of the observed outcomes.

6.2 Changing our attention layer

Following the failure of our initial hypothesis, we aimed to assess the impact of two alternative approaches on the model's performance. Specifically, we sought to compare the outcomes obtained by appending each token from the BERT model to a sentiment BERT token with those obtained by solely using the class tokens from both the distill BERT and sentiment BERT models. Our objective was to determine the effect of these modifications on the model's overall results.

And Here are the results:

Size of Training Data	F-1 score	Precision	Recall	Accuracy
10 percent	0.85	0.84	0.86	0.84
25 percent	0.86	0.86	0.86	0.86
50 percent	0.88	0.86	0.90	0.88
100 percent	0.90	0.91	0.90	0.90

7 Conclusion

Through our investigation, we learned that incorporating knowledge from a sentiment analysis model into a mental health classification model may not always lead to improved performance. Despite our hypothesis that feature vectors derived from sentiment analysis could aid in the classification of mental health, our results showed that the sentiment analysis-assisted model did not consistently outperform the baseline model. Furthermore, it is possible that any perceived performance gain may be attributed more so to the attention-mechanism rather than the sentiment analysis assisted model. We also learned that for this particular task, relatively good performance can still be achieved through transfer learning with a BERT model and only using a few 1000 samples for training.

References

Hyeju Jang, Seo-Jin Bang, Wen Xiao, Giuseppe Carenini, Raymond Ng, and Young ji Lee. 2021. Kw-attn: Knowledge infused attention for accurate and interpretable text classification. *Workshop on Knowledge Extraction and Integration for Deep Learning Architectures; Deep Learning Inside Out.*

Annavarapu Chandra Sekhara Rao Mayur Wankhade. 2022. Opinion analysis and aspect understanding during covid-19 pandemic using bert-bi-lstm ensemble method. *Scientific Reports 12 Article number: 17095.*

Mohammed A. Alqarni Rukhma Qasim, Waqas Haider Bangyal and Abdulwahab Ali Almazroi. 2022. A fine-tuned bert-based transfer learning approach for text classification. *Hindawi: Journal of Healthcare Engineering Vol. 2022.*

Adarsh S and Betina Antony. 2022. Transfer learning using bert for detecting signs of depression from social media texts. *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion.*