

The Effect of Using Sentiment Analysis Features on Mental Health Classification

Michal Lange, Thushara Manjari Naduvilakandy,
Fagun Raithatha, Daniel Stapleton



Task Description

- **Research Problem**
 - The focal point of our research inquiry is whether the incorporation of a knowledge-enriched attention layer could enhance the accuracy of mental health classification, even when using a limited dataset.
- **Reason behind our decision to opt for this particular task**
 - Early detection of mental health issues can expedite treatment .
 - Sentiment analysis involves classifying sentiment in a given document, which can involve positive and negative emotions on a given subject. Our observations from the mental health dataset suggests that the expression of positive or negative emotion may be conducive to classification.



Objectives

- The objective of our investigation is to determine whether the integration of knowledge derived from Sentiment Analysis would improve the performance of the mental health classification model.
- We have a DistilBERT model trained on twitter data for Sentiment Analysis, which is similar in style, speech and length to our own dataset.
- Compare the effectiveness of a baseline classification model versus sentiment analysis assisted classification model and monitor performance fall-off with smaller dataset sizes.



Prior Work

- We have found a few articles that touch on the idea we are proposing. Unlike our idea, these papers do not combine BERT output tokens with sentiment analysis vectors as inputs into a mental health classification model.
- The first paper that we found relevant to our proposal is about Multi-Feature Fusion We can use the idea of feature fusion for combining the BERT output with the Sentiment Analysis output for our experiments.

Eke CI, Norman AA, Shuib L (2021) Multi-feature fusion framework for sarcasm identification on twitter data: A machine learning based approach. PLOS ONE 16(6): e0252918. <https://doi.org/10.1371/journal.pone.0252918>

- The second paper that we believe we can draw inspiration from is about using BERT with knowledge distillation and a Bi-LSTM to analyze twitter and Reddit posts for depression and anxiety signs.

Zeberga K, Attique M, Shah B, Ali F, Jembre YZ, Chung TS. A Novel Text Mining Approach for Mental Health Prediction Using Bi-LSTM and BERT Model. Comput Intell Neurosci. 2022 Mar 3;2022:7893775. doi: 10.1155/2022/7893775. PMID: 35281185; PMCID: PMC8913054. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8913054/#sec3title>

- The third paper we have found to be useful is a paper titled KW-ATTN: Knowledge Infused Attention for Accurate and Interpretable Text Classification written in part by Dr. Jang. We have turned to this paper to help solve our attention layer concatenation of Bert tokens and/or cls tokens.

Hyeju Jang, Seojin Bang, Wen Xiao, Giuseppe Carenini, Raymond Ng, and Young ji Lee. 2021. KW-ATTN: Knowledge Infused Attention for Accurate and Interpretable Text Classification. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 96–107, Online. Association for Computational Linguistics. <https://aclanthology.org/2021.deelio-1.10/>

- We have also found some help dealing with the attention layer of our project. We have looked into multiple source (Google, Stackoverflow, etc) but found this article on Kaggle that dealt with attention layers to help us develop that part. <https://www.kaggle.com/code/mlwhiz/attention-pytorch-and-keras>



Dataset Description

- 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**.
- Link : <https://www.kaggle.com/datasets/reihanenamdari/mental-health-corpus>
- The mental health dataset has an equal distribution of cases, with 51% 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% are labeled as 0 which means that there were no flags raised by the verbiage used in the post.
- Reihaneh Namdari has uploaded the dataset, but no information has been provided regarding the dataset's specifics or its acquisition process.
- On average, there are 72 words present in the comments section of the data.



Dataset Description

A post that has been categorized as 0, meaning the post contains inflammatory words that could be considered toxic or negative in nature.

nothing look forward life i dont many reasons keep going feel like nothing keeps going next day makes want hang myself

A post that has been categorized as 1, meaning the post had no indications of a mental health issue.

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

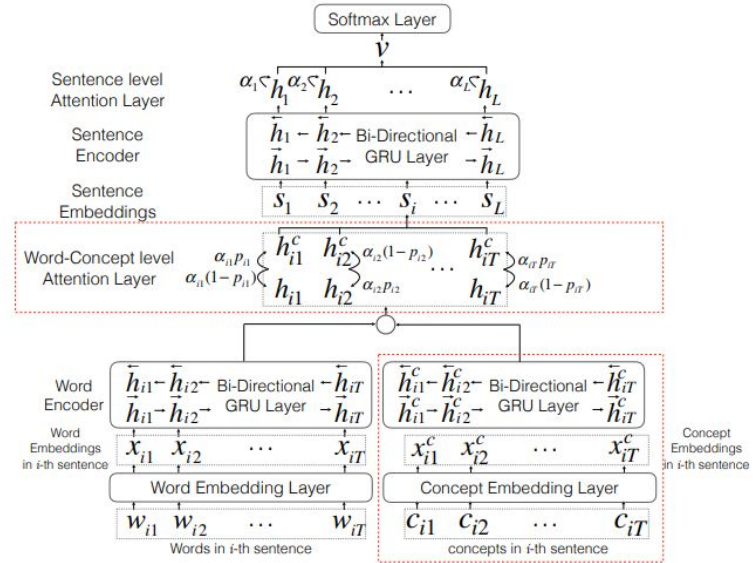


Model Proposal

- Baseline Model :
 - DistilBertForClassification (Base)
 - DistilBertForClassificationSentimentAnalysis (Fine-tuned for twitter sentiment analysis task)
 - Similarity in style of text used for SA task and our task
- Our Model
 - Incorporate both of the above models into a single model and combine outputs using attention-based mechanism.
- Our Model Version 2.
 - Implement attention-based mechanism but only for the base DistilBERT model.

Word-Concept Attention

$$\begin{aligned}
 u_{it} &= \tanh(W_\alpha[h_{it}, h_{it}^c] + b_\alpha) \\
 p_{it} &= \text{sigmoid}(w_p[h_{it}, h_{it}^c] + b_p) \\
 \alpha_{it} &= \frac{\exp(u_{it}^T u_\alpha)}{\sum_t \exp(u_{it}^T u_\alpha)} \\
 s_i &= \sum_t \alpha_{it} ((1 - p_{it})h_{it} + p_{it}h_{it}^c)
 \end{aligned}$$



Attention Layer

```
def forward(self, outputs_base, outputs_sentiment, attention_mask=None):
    # Concatenate the outputs of both models
    hidden_concat = torch.cat((outputs_base, outputs_sentiment), dim=2)
    # Apply cross-token attention
    cross_token = torch.tanh(self.cross_token_param(hidden_concat))
    # Apply individual token attention
    ind_token = torch.sigmoid(self.ind_token_param(hidden_concat))
    # Apply attention weights
    att_token = torch.squeeze(torch.matmul(cross_token, self.att_token_param))
    # Apply softmax to get attention weights and apply attention mask
    att_token = self.masked_softmax(att_token, attention_mask, 1)
    # Expand attention weights to match the hidden states dimension
    att_token = att_token.unsqueeze(2)
    att_token = att_token.expand(-1, -1, 768)
    # Get the inter token representation
    inter_token = (1 - ind_token) * outputs_base + ind_token * outputs_sentiment
    # Apply attention weights to the hidden states
    final_representation = torch.mul(att_token, inter_token).sum(dim=1)

    return final_representation
```



Baseline Model Results

Using : distillbert-base-uncased

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

Using : 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



Results from our Model

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

Results: Cross-Dim Tuning

Fine-Tuning model with different Cross Dimension


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

Testing Different Data size on best Cross-Dimension - 568

Training Data Size	F-1 Score
0.1	0.89
0.5	0.87
1	0.94

Training Data Size	F-1 Score
1	0.94
1	0.90
1	0.93

Standard Output : 0.923



Is the two model system responsible for the improved result?

$$u_{it} = \tanh(W_{\alpha}[h_{it}, \cancel{h_{it}^c}] + b_{\alpha})$$

$$\cancel{p_{it}} = \cancel{\text{sigmoid}(w_p[h_{it}, h_{it}^c] + b_p)}$$

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

$$s_i = \sum_t \alpha_{it} ((1 - \cancel{p_{it}})h_{it} + \cancel{p_{it}}h_{it}^c)$$

Accuracy: 0.9056468906361687, F1: 0.9104477611940298, Precision: 0.8742671009771987, Recall: 0.9497523000707714



What is left before the Report

- Collect data from repeated training runs
- Direct comparison of the performance of both single and dual, attention-based models.
- (If time is left) compare the concatenation of the class tokens of both models for classification.



Lessons Learned

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.



Lessons Learned

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.