DeepBlues@LT-EDI-ACL2022: Depression level detection modelling through domain specific BERT and short text depression classifiers

Nawshad Farruque, Osmar R. Zaïane, Randy Goebel
Alberta Machine Intelligence Institute
Department of Computing Science
Faculty of Science
University of Alberta, Edmonton, AB, Canada T6G 2E8
{nawshad,zaiane,rgoebel}@ualberta.ca

Sudhakar Sivapalan
Department of Psychiatry
Faculty of Medicine and Dentistry
University of Alberta, Edmonton, AB, Canada T6G 2B7
sivapala@ualberta.ca

Abstract

We discuss a variety of approaches for building a robust depression level detection model from longer social media posts (e.g., Reddit depression forum posts) using a mental health text informed pre-trained BERT model. Further, we report our experimental results based on a strategy to select excerpts from long text and then fine-tune the BERT model to combat the issue of memory constraints while processing such texts. We show that, with domain specific BERT, we can achieve reasonable accuracy with fixed text size (in this case 200 tokens). In addition we can use short text classifiers to extract relevant text from the long text and achieve some accuracy improvement, albeit, trading off with the processing time for extracting such excerpts.

1 Introduction

Depression has been found to be a major cause behind at least 800,000 deaths committed through suicide each year worldwide[1]. Moreover, It has been found in earlier research that depressed individuals show help seeking behavior through their social media posts (Guntuku et al., 2017). So analyzing social media posts for depression detection is an important research area (Coppersmith et al., 2014; Mowery et al., 2017). In our work, we analyze Reddit social media posts to identify whether a particular post exhibits either of three levels of depression, including (1) No Depression, (2) Moderate Depression, and (3) Severe Depression, as a part of a shared task challenge (Sampath et al., 2022). We use a state-of-the-art transformer-based model called BERT, which was pre-trained on mental health related social media data. Further, we compare our model with two variations of the same, with other models trained on (1) relevant excerpt extracted Reddit posts and (2) a subset of depressive sentences in Reddit posts calculated with the help of short text classifiers. In the next sections, we elaborate on each of our strategies.

2 Depression Level Detection through Fine-tuning Mental BERT (MBERT)

BERT (Devlin et al., 2018), which stands for Bidirectional Encoder Representations from Transformers, has been found to be very effective in different downstream NLP tasks such as, text classification (Sun et al., 2019) and Depressive post detection (Ji et al., 2021). Here we use a mental health pre-trained BERT model, called Mental BERT (MBERT), which was pre-trained on several mental health forums under Reddit (Ji et al., 2021). Further, we fine-tune this model on the provided training dataset (Kayalvizhi and Thenmozhi, 2022) for this shared task. Since fine tuning BERT based models on longer text requires significant memory resources, we limit our text data to the first 200 tokens, which covers around 70% of the total samples provided. Before feeding input to our model, we convert all texts to lower case, and use

[1] https://who.int/mental_health/prevention/suicide/suicideprevent/en/

Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 167 - 171

an uncased version of the MBERT model for fine tuning. We also experiment with further enhancement of our classifier by fine tuning it through a selection of 200 "relevant" tokens from constituent Depressive sentences for a post from the training sample, which are longer than 200 tokens and also depression-indicative. In addition, we investigate whether the distribution of constituent depressive sentences in each posts also have some predictive power for this task.

3 Extracting Relevant Excerpts for Fine-tuning Mental BERT (RE-MBERT)

To extract relevant excerpts, we use a majority voting classifier (MVC) which is built using four depressive short text or Tweet classifiers. Three of these classifiers use different pre-trained word and sentence embeddings and represent each sentence through either averaged embedding of all the constituent words of that sentence, or the sentence embedding of the sentence itself. The left classifier uses Zero-shot modelling for classifying each sentence for signs of depression. Description of these classifiers including the datasets they were trained on, have been previously described (Farruque et al., 2019, 2021). Since we cannot extract more than 200 tokens for each of our posts and, within those 200 tokens, we may not have all the relevant tokens which are important for this task, we plan to extract relevant (or depressive) constituent sentences or excerpts from the posts which have more than 200 tokens and which are labeled as either carrying signs of "Moderate" or "Severe" depression in the training set. To do this, we parse each post by exploiting punctuation, i.e. ".", "?" and "!" to find its constituent sentences. We then feed those sentences to MVC, where the above mentioned four short text classifiers vote to indicate whether the constituent sentences of a post are depression-indicative or not; we only take a sentence as a representative text for depression if at-least three of those four classifiers agree. We apply the same short text pre-processing as we did while training our short text classifiers to clean each sentences within our posts. In this cleaning process, with the help of a python library named "Ekphrasis" (Baziotis et al., 2017), we re-contract word contractions, replace elongated words in their original form, convert all to lower case and remove non-words, so our cleaned sentence is mostly regular words separated by spaces. Finally after fine-tuning our MBERT classifier (see Figure 1), we infer the labels from the provided test set and use extracted excerpts only for the posts having greater than 200 tokens (see Figure 2). In summary, in our training set, we only extract excerpts when a post is depression-indicative and longer than 200 tokens. If no excerpts are extracted or the post is less than 200 tokens long or it is not depression-indicative, then we use the cleaned version of the original post and feed it to our MBERT classifier. After this procedure is completed, the total posts left beyond 200 tokens were 1667 which is more than a 50% reduction than the original number of posts having more than 200 tokens (i.e., 4018). All the posts from the original training set beyond 200 tokens and with depression indication are now pre-processed so that those now contain only depressive excerpts which is important for our classification. Our assumption with the posts with less than or equal 200 posts is that, they have more depressive sentence density than their longer counter-parts.

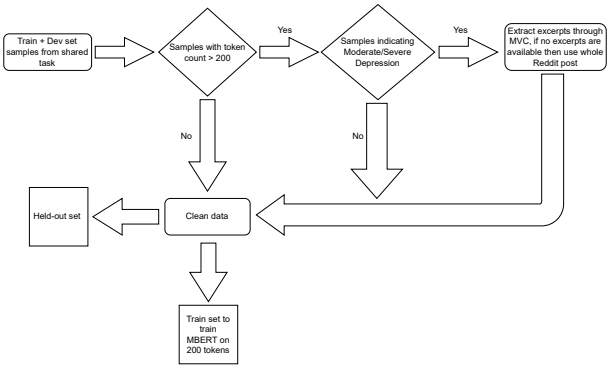

Figure 1: RE-MBERT training/fine-tuning algorithm

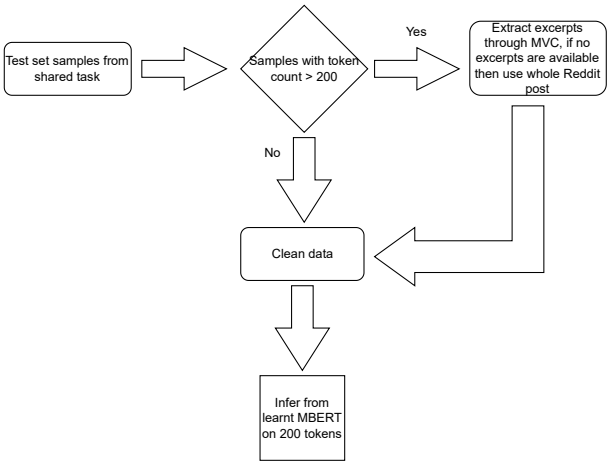

Figure 2: Fine-tuned RE-MBERT testing/inference algorithm

3.1 Depressive Sentence Proportion based Method (SPROP)

We calculate the number of sentences which are depression-indicative out of total number of sentences in a post: we call this the Depressive Sentence Proportion Value (DSPV). In our training set, we calculate DSPV for our depressive posts and we assume this to be 0 for the "No Depression" class, as ideally there would not be any depression-indicative sentences or might be too few such sentences present in this class. Later we use this as a feature extracted from all our training samples and train our model and report result in test samples based on same extracted feature.

4 Experimental Setup

We mix the training and development set provided in the shared task data, which is a total of 13,387 samples, and then split it into a training set of size: 12,589, and validation set of size: 128, and held-out set of size: 670 samples. For the MBERT classifier we use uncased mental BERT [2]. We use maximum token size = 200, number of epochs = 10, training and test batch size = 16. We employ a NVIDIA-GeForce-RTX 3070 GPU with 8 GB of integrated memory and 32GB of RAM.

For SPROP, we use a MLP classifier [3] with default settings and max iteration value of 300, during label inference time we take the argmax of the output label probabilities using $predict\_proba()$ function.

Although BERT-based modelling takes around 30 minutes for training and testing, excerpt extraction for creating RE-MBERT takes a number of days to complete.

In the next section we report and analyze the performance of our top models, i.e. MBERT and RE-MBERT.

5 Result Analysis

We report both label based accuracy for our held-out set over all samples (see Tables 1 and 2) and overall accuracy scores (i.e., Avg. Precision, Recall, Weighted-F1 and Macro-F1 across all labels over all samples, see Table 3) for our held-out set. Also, we report overall accuracy scores (i.e., Avg. Precision, Recall and Weighted F1 and Macro-F1 over all labels across all samples) for test set provided by the shared task organizers (see Table 4). From the F1-scores for different labels in the held-out set, we see added value to the classification performance for "No Depression" and "Severe Depression" classes (see Tables 1 and 2). For "Moderate Depression" class there is still some improvement but it is not very pronounced (only 0.26%). Increased recall in "No Depression" and "Severe Depression" classes (by almost 4.5% and 8%) indicates that the classifier learns a high false positive rate or is more inclined to erroneously identify a post as either not depressive or severely depressive through our training procedure. However, for our "Severe Depression" class, our classifier also achieves better precision scores, means robustness against false negative results. For "Moderate Depression" class we also see 1.6% precision improvement but with a cost of 1.73% decrease in recall. We can see the reflection of these results in Table 3, with RE-MBERT having significantly better Macro-F1 score due to pronounced recall for "No Depression" (by 3.14%) and "Severe Depression" (by almost 6%) classes.

Classes	Precision	Recall	F1-score
No Depression	**0.7812**	0.4975	0.6079
Moderate	0.7371	**0.9113**	**0.8150**
Severe	0.5500	0.3492	0.4272

Table 1: Accuracy scores for each labels for held-out set on MBERT

Classes	Precision	Recall	F1-score
No Depression	**0.7786**	0.5423	0.6393
Moderate	0.7531	**0.8941**	**0.8176**
Severe	0.5625	0.4286	0.4865

Table 2: Accuracy scores for each labels for held-out set on RE-MBERT

In the test set accuracy scores in Table 4, we see our strategy (RE-MBERT) helps in achieving a slightly better Macro-F1 score (by 0.3%) whereas the precision score improvement is more pronounced (by 1.8%) than recall compared to MBERT. Additionally, improvement in the Weighted-F1 score (by 1%) suggests that our strategy helps improve the F1-score for one of our Depression level classes. Unfortunately, since we do not have access to test set labels we cannot do detailed label-wise error analysis. We also test

[2] https://huggingface.co/mental/mental-bert-base-uncased

[3] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

Experiment Name	Recall	Precision	Weighted-F1	Macro-F1
MBERT	0.5860	0.6894	0.7164	0.6167
RE-MBERT	**0.6216**	**0.6981**	**0.7330**	**0.6478**

Table 3: Avg. accuracy scores for held-out set across all labels over all samples

Experiment Name	Recall	Precision	Weighted-F1	Macro-F1
MBERT	**0.5431**	0.5374	0.6442	0.5374
RE-MBERT	0.5345	**0.5554**	**0.6542**	**0.5404**
OPI (top model)	0.5912	0.5860	0.6660	0.5830

Table 4: Avg. accuracy scores for test set across all labels over all samples

with a single feature SPROP method, which results in Macro-F1 score of 0.3387 in test set. We found SPROP is more robust for more populated classes such as "Moderate Depression" and "No Depression" and performs poorly for the "Severe Depression" class. This seems reasonable because a single feature has less predictive value. We tried that method just to observe whether depressive sentence proportion as a feature has any significance or not. In future, we would like to use this with other features in future to make our modelling robust.

Finally, our MBERT modelling does not perform data cleaning as RE-MBERT and SPROP. We find that data cleaning does not provide any significant performance gain, by comparing MBERT trained with cleaned and not cleaned samples. With data cleaning, we achieve only 0.48% accuracy gain in the held-out set. Therefore we believe that the accuracy increase in the held-out and test set for our RE-MBERT modelling is purely attributed to our excerpt extraction algorithm. The Held-out set sample distribution is similar to our training set, which explains why we have better accuracy scores there.

6 Conclusion

We have described a few strategies for the Depression level detection shared task from Reddit posts. We use state-of-the-art mental health data pre-trained BERT model (MBERT) and further fine-tune it with the shared task data and achieve 7th position in terms of Macro-F1 score and 3rd position in terms of Weighted F1 score compared to 30 other participating teams. We also present a strategy (RE-MBERT) to consider while training MBERT in a resource constrained environment through a subset of relevant sentence selection for longer posts. Our strategy shows some improvements in both training and test set which is stimulating and encouraging.

Acknowledgements

We are grateful to the Alberta Machine Intelligence Institute (AMII) for their generous funding for our research.

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.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60.

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*.

Nawshad Farruque, Randy Goebel, Osmar R Zaïane, and Sudhakar Sivapalan. 2021. Explainable zero-shot modelling of clinical depression symptoms from text. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1472–1477. IEEE.

Nawshad Farruque, Osmar Zaiane, and Randy Goebel. 2019. Augmenting semantic representation of depressive language: From forums to microblogs. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 359–375. Springer.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on

social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. Mentalbert: Publicly available pretrained language models for mental healthcare. *arXiv preprint arXiv:2110.15621*.

S Kayalvizhi and D Thenmozhi. 2022. Data set creation and empirical analysis for detecting signs of depression from social media postings. *arXiv preprint arXiv:2202.03047*.

Danielle Mowery, Hilary Smith, Tyler Cheney, Greg Stoddard, Glen Coppersmith, Craig Bryan, and Mike Conway. 2017. Understanding depressive symptoms and psychosocial stressors on twitter: a corpus-based study. *Journal of medical Internet research*, 19(2).

Kayalvizhi Sampath, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, and Jerin Mahibha C. 2022. Findings of the shared task on Detecting Signs of Depression from Social Media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China national conference on Chinese computational linguistics*, pages 194–206. Springer.