ICIC Express Letters
Part B: Applications
Volume **13**, Number **9**, September **2022**

ICIC International ©2022 ISSN 2185-2766
pp. 983–989

DEPRESSION DETECTION ON REDDIT FORUM POSTS USING BERT & DEBERTA

Hansen Riady Kwee* and Amalia Zahra

Computer Science Department, BINUS Graduate Program, Master of Computer Science
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
amalia.zahra@binus.edu
*Corresponding author: hansen.kwee@binus.ac.id

Received November 2021; accepted February 2022

Abstract. *Depression is one of the mental disorders suffered by many people in the world. People usually keep it to themselves and do not seek treatment. Proper treatment is needed to prevent people from suicide. To know how people with depression behave, we can diagnose them from the language that they used every day. Because the Internet is one of the most used services in the world, we can learn from the text they posted on social media. In this work, we used Reddit posts as the main data source to learn about depressive text. We then finetuned BERT and DeBERTa models with the collected dataset. Using this method, the highest accuracy achieved is 96.76% using the BERT model. From our experimental results, BERT and DeBERTa are able to learn about depressive and non-depressive text.*

Keywords: Depression, Depression detection, Transfer learning, Transformer model, Reddit, BERT, DeBERTa

1. Introduction.

Depression is one of the common mental disorders that is suffered by people in the world with 264 million people [1], as illustrated in Figure 1 which shows the number of people who suffered from mental disorders. While there are so many people who experienced depression, the research found that not many people seek treatment when they feel depressed. There are only about 10%-20% of people who seek treatment in low to middle-income countries [2]. Without a proper diagnosis and treatment, depression can lead people to suicide.

Social media has become part of our life and is often a place for people with mental disorders to express their feeling and seek support [3]. Therefore, social media can be used as a place to learn whether someone is depressed or not. People suffering from depression tend to use certain words to indicate that they are depressed. Research also suggested that content from online communities can be utilized to help detect mental health problems [4]. Calvo et al. explained how data from social media is used to learn about people's mental health conditions by using NLP techniques [5]. They found that people who likely suffer from mental health tend to act differently on social media. Several social media platforms have been used to detect mental disorders, such as Reddit. Reddit is a social media platform with many communities discussing different topics called subreddits [6]. Every subreddit has its forum discussing the related topic. An example of a subreddit is r/earth which is a forum discussing planet earth.

Reddit has been used in research to detect mental disorders. Research is conducted by Shen and Rudzicz in detecting anxiety disorders on Reddit [7]. The study used post data from the r/anxiety, r/healthanxiety, r/socialanxiety, r/panicparty subreddits as data categorized as anxiety disorder and that from several other subreddits such as r/askscience,

DOI: 10.24507/icicelb.13.09.983

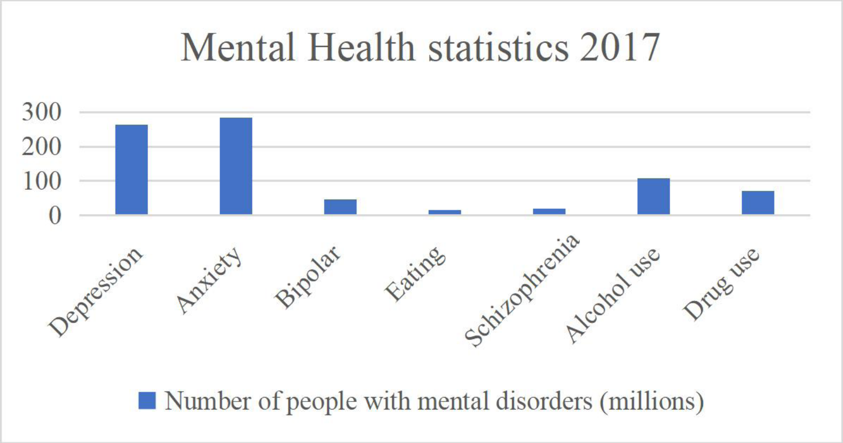

Figure 1. Mental health statistics in 2017

r/writing, and r/books as data categorized as no anxiety disorder. The study used the vector space embeddings method: Word2Vec and Doc2Vec, LDA topic modeling, LIWC features, and N-gram language models to categorize words frequently used by people with anxiety disorders and those without anxiety disorders. Research on detecting depression is also conducted by Pirina and Çöltekin in identifying depression on posts from several social media, especially Reddit. They experimented with several datasets collected from depression support forums, non-depression support forums, depression subreddit, breast cancer subreddit, subreddit related to family and friends, posts with depression posted on other forums, and posts with no depression posted on other forums [8]. The study is able to get a higher F1 score from the dataset collected from Reddit. This method of using Reddit as a data source for depression detection has been further improved by Tadesse et al. by using several feature extraction on the data to gain words related to depression before using it as input to the classifier model [9]. They found that using feature extraction is effective for the models to learn linguistic words that depressed people used. Although they are able to achieve good results, there is still room for improvement in terms of good dataset and performance. To achieve a better performance, different models are used. Chen et al. showed that the BERT model can achieve high accuracy in recognizing text [10]. In this work, we attempt to improve the performance from the previous work on detecting depression on Reddit social media by using a new dataset collected from several subreddit and finetune BERT and DeBERTa from transformers for the classification. BERT and DeBERTa are used in this experiment because both of the models are able to achieve state-of-the-art results in the NLP problem [11].

In summary, this research had been expected to improve the accuracy of depression detection from Reddit posts by using new data collected from Reddit and the built-in feature extraction and models from the transformer. By performing these methods, we were able to improve the performance of depression detection to 96.76% of accuracy.

In Section 2, we examine the problems related to our research. We then explained our step of the experiment in Section 3. The results and discussions of our experiment are shown in Section 4. Finally, our conclusions are presented in Section 5.

2. Problem Statement and Preliminaries.

Mental health has become a serious problem in our society. People who suffer from mental disorders tend to use more social media than talk to real people [12]. Identifying specific mental health disorders from text is not an easy task, especially from social media. Research from Calvo et al. showed that there are specific characteristics on how people with mental health disorders behave in social media [5]. There have been many researches on detecting mental disorders from social media posts especially Reddit [7-9,13,14]. Reddit is used because it is one of the social media platforms that becomes the place for people with mental health to share their stories. One such research was conducted by Shen and Rudzicz on detecting anxiety from posts on Reddit [7]. They collect the data from several subreddits related to anxiety and non-anxiety. Feature extraction is used to get the topics related to anxiety. They used Word2Vec and Doc2Vec, LDA topic modeling, LIWC features, and N-gram language model to categorize the topics. The feature is then classified with Support Vector Machine (SVM), Linear Regression (LR), and Neural Network (NN). They can achieve an accuracy of 99% with LIWC features combined with NN. Detecting other mental health disorders like bipolar is also conducted by Sekulic et al. from Reddit [13]. This study uses a dataset consisting of several topics where each topic contains posts from users who are categorized as having bipolar and those who are categorized as not having bipolar. The researchers used the SVM, LR, and Random Forest (RF) for the classification. They were able to achieve an accuracy of 86%. The researchers suggest that people with bipolar disorder tend to use first-person pronouns.

For depression detection, research was conducted by Pirina and Çöltekin by using posts from several social media platforms about depression and non-depression which are depression support forums, non-depression support forums, depression subreddit, breast cancer subreddit, subreddit related to family and friends, posts with depression posted on other forums, and posts with no depression posted on other forums. They used SVM as the classifier and found that with a good selection of data, the model was able to achieve a good result in detecting depression [8]. This research was further improved by Tadesse et al. by adding feature extraction which was N-grams, LIWC, LDA to extract the feature from the datasets. The feature was then trained with LR, SVM, RF, Adaptive Boosting (AdaBoost), and Multilayer Perceptron (MLP). They were able to achieve a good result by combining all the features [9]. From those studies, we can see that using datasets related to depression and combining feature extraction can identify depression from the text. Another approach on depression detection was conducted in early detection of depression in CLEF eRisk 2017 by Shah et al. [14]. They used a total of five features which were TrainableEmbed, GloveEmbed, Word2VecEmbed, FastextEmbed, and Metadata features. The features were then fed into the Bidirectional Long Short-Term Memory (BiLSTM). The output from BiLSTM was fed into the first hidden layer alongside the feature from meta-features into the second hidden layer. They were able to achieve good performance by combining Word2VecEmbed and Meta features [14]. Another research on depression detection was carried out by Orabi et al. on Twitter [15]. They used word embeddings as the feature from tweets. They compared the performances from some models, which were Convolutional Neural Network (CNN) and BiLSTM to detect depressive text. They were able to achieve an accuracy of 87.96% [15]. Another review on depression detection methods was conducted by William and Suhartono [16]. They found that BiLSTM is the current best method to detect depression text. They also experimented with BERT to detect depression and obtained an accuracy of 92%. To further learn about the characteristics of depressive text, we can improve the performance of identifying depression by more data and new state-of-the-art NLP models. These methods are conducted to better understand the relationship between a depressed person and the word they used on the Internet. In our research, we try to use state-of-the-art models from transformers to obtain a better performance in identifying depression.

3. Methodology.

Based on the previous work and the models explained in the previous section, we performed our research according to the diagram shown in Figure 2. Each of the experiment steps is explained in the following section.

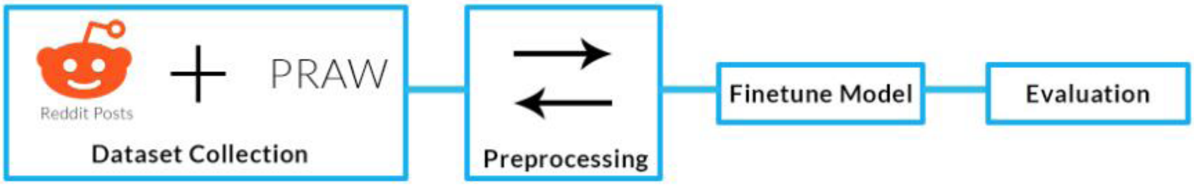

FIGURE 2. Experimental steps

3.1. **Dataset collection.** Reddit is used as our main source of data. We collected a total of 4400 data samples from several subreddit. The dataset is collected using *Python Reddit API Wrapper* (PRAW). There are a total of 6 subreddits that are used to collect the data: 3 subreddits for depression-related and 3 subreddits for non-depression-related. The list of the subreddits is shown in Table 1. The dataset is labeled based on the source of the subreddits. Label 1 is given to data from depression-related and label 0 is given to that from non-depression-related.

TABLE 1. List of subreddits that is used

Subreddit	**Category**	**Label**
r/depression	Depression	1
r/depressed	Depression	1
r/depression_help	Depression	1
r/askscience	Control	0
r/books	Control	0
r/writing	Control	0

3.2. **Dataset preprocessing.** The dataset collected is then processed before being used for training. There are 3 steps of preprocessing that is performed to this dataset.

1) Transform Cases, each letter in the dataset is converted to lowercase. It aims to improve the consistency of the dataset.
2) Text Normalization, all the words that contain abbreviation, misspelling, numbers combined with letters will be transformed into their normal forms.
3) Tokenization, the last step of the preprocessing where the tokens [CLS], [SEP], and [PAD] are added to every text before being fed into the transformer layers.

The dataset is then split into two parts, 80% for training and 20% for testing. The training portion is then split into training and validation sets (75% for training and 25% for validation).

3.3. **Finetune model.** For our experiment, we mainly used two models from transformers to classify depression. We used BERT and DeBERTa base for the models.

A. **BERT** or Bidirectional Encoder Representations from Transformer is a model based on transformer architecture that consists of a multilayer of encoder. BERT used the attention mechanism to learn the information between words. BERT is pre-trained with 2 tasks before we can use it for other tasks. The pre-training process is trained with the dataset from BooksCorpus and English Wikipedia on masked language model and next sentence prediction. To use BERT on another task, we need to add an output layer according to the tasks. By applying transfer learning to this model, we were able to achieve state-of-the-art performance on several tasks [11].

B. **DeBERTa** or Decoding-enhanced BERT with disentangled attention is a model with a new architecture based on BERT and RoBERTa. DeBERTa introduces two new methods which are disentangled attention mechanism and enhanced mask decoder.

Disentangled attention changes the input representation as two vectors which consist of the word and the position of the word. Enhanced mask decoder enables the model to get the absolute position of every word at the end of the transformer layer. This absolute position can help predict the word [17].

Before the text was fed into the models, we encoded the text to get three layers, which were input_ids, attention_mask, and token_type_ids. These three layers were then used as the inputs to the models. For every connected layer on the models, we added a dropout with the value of 0.2. We took the CLS output from the models and fed it into a dense layer. The dense layer was calculated with the sigmoid activation function. The details of the architecture can be seen in Figure 3.

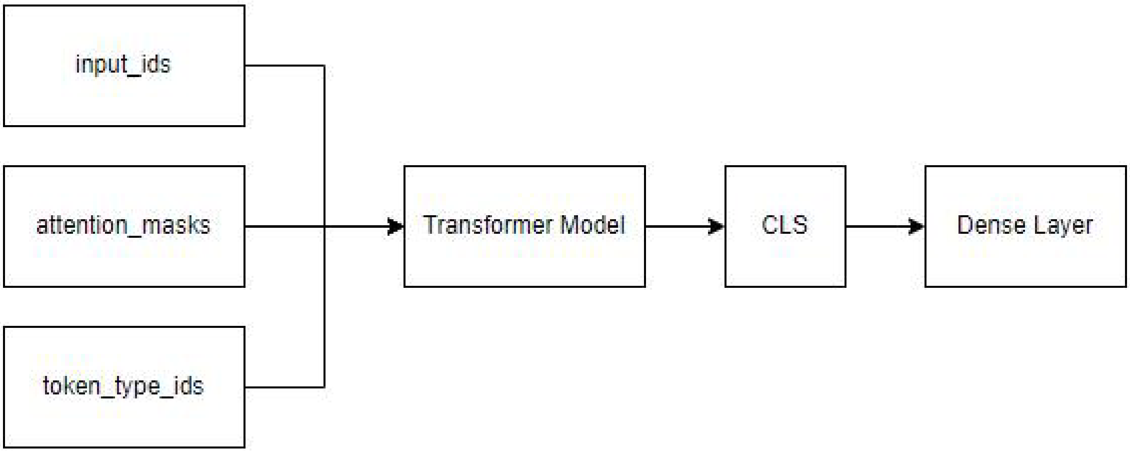

Figure 3. Model architecture

The finetune phase was carried out on Google Colab Notebook. The resource used in the training process was GPU. For the models, we used the Tensorflow version provided by huggingface[1]. The finetune phase was carried out to make the model learn about depressive text by finding the most suitable learning rate and epoch for both models.

3.4. **Evaluation.** The model was finetuned until a good learning rate and epoch for each model were found. After the training, we evaluated the performance of the models with the confusion matrix method. We compared the value of the F1 score, accuracy, precision, and recall from each model. We also calculated the loss from every epoch of the training by using the binary cross-entropy method.

4. **Results and Discussion.** Our main goal in this research is to find the best model to detect depression from Reddit posts. In our experiment, we first try different learning rates. We found good results from two learning rates, which are 2e-6 and 1e-6. With the selected learning rates, we trained the models and found that both models reached convergence in different epochs. We decided to use a total of 10 epochs during the training to observe the convergence. The batch size used in the training was 32. The max sequence length used for the transformer model was 128. We then used the models to evaluate our test set. Both learning rates show good classification results on overall metrics.

4.1. **Validation.** After finetuning for a total of 10 epochs and 3520 rows of data, we observed the loss from both models to gain insights on when the model reached convergence. We observed that both models reached convergence on different epochs according to the different predefined learning rates. For BERT, the model reached convergence on epoch 2 with the learning rate of 2e-6 and epoch 5 with the learning rate of 1e-6. DeBERTa model reached convergence on epoch 6 with the learning rate of 2e-6 and epoch

[1] https://huggingface.co/models

8 with the learning rate of 1e-6. We then recorded the results based on the epoch when the models reached convergence, as we stated before. The validation results are shown in Table 2. From the results, BERT showed a higher performance across all metrics except recall compared to DeBERTa. DeBERTa is able to achieve a higher recall than the BERT model which means that the model had a more correct prediction on depression text. With a higher learning rate, both models can reach convergence faster and yield a better performance result.

Table 2. Validation results for both learning rates

Model	**lr = 2e-6**				**lr = 1e-6**			
	Accuracy	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**
BERT-base	98.03%	99.20%	97.62%	98.40%	97.45%	99.19%	96.44%	97.80%
DeBERTa-base	96.88%	96.86%	97.82%	97.34%	95.60%	97.36%	95.04%	96.18%

4.2. **Testing.** After obtaining the validation results, we save the model's weights from the epoch stated above. We load the saved weights into the models. The model is then used to evaluate our testing data with a total of 880 rows of data. The testing results are shown in Table 3. From the testing results, BERT still shows higher performance across all metrics except recall on learning rate of 2e-6. Although the accuracy produced by the DeBERTa model is lower than BERT, the difference is insignificant.

Table 3. Testing results for both learning rates

Model	**lr = 2e-6**				**lr = 1e-6**			
	Accuracy	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**
BERT-base	96.76%	98.77%	95.63%	97.17%	97.45%	99.19%	96.44%	97.80%
DeBERTa-base	96.30%	96.67%	97.05%	96.86%	95.49%	96.23%	96.04%	96.13%

From the validation and testing results stated above, we can see that both models performed well on detecting depression from Reddit posts. Both can achieve average score of 95+% across all metrics from the confusion matrix. The best model is achieved by BERT. Although DeBERTa is an upgrade from BERT, BERT still shows better results in detecting depression. This performance result is also achieved not because of the models but rather the data collected from Reddit. With a total of 4400 rows of data, the transformer models are able to discriminate depressive and non-depressive text. Compared to other works related to depression detection [9,14], the models achieved an accuracy of 91% and 81%, respectively with the combination of feature extraction. Our models still achieve better results despite the difference in the dataset. We found that built-in feature extraction on BERT which is Wordpiece Embeddings is better for the model to learn about depressive text than using another feature extraction. Our results are also better in maintaining performance on both validation and testing results compared to [16] with an accuracy of 92% in using BERT models.

5. **Conclusions.** In this paper, we tried to get the best model to detect depression from the text. We performed it by using text classification techniques and transformer models. Based on our experiments, we compared two models, which were BERT and DeBERTa, for this task and obtained good performance results. To make the models learn about the depressive text, we collected the data from Reddit. We then finetuned both models with our data that has been processed with several learning rates. Our finetuned models are able to achieve high performance across all metrics. From all the models that we applied, BERT is able to achieve the highest performance. From this experiment, we can conclude

that we were able to detect depression from Reddit posts by training them with depressive and non-depressive data.

For future work, external assistance from experts is necessary to help annotate the data with depressive or non-depressive label. Such an effort might improve the model performance. In terms of the advantage offered by this research, the model can be implemented in healthcare to further automate the process of detecting people with mental health problems more early.

REFERENCES

[1] H. Ritchie and M. Roser, *Mental Health*, https://ourworldindata.org/mental-health, 2018.

[2] S. Evans-Lacko, S. Aguilar-Gaxiola, A. Al-Hamzawi, J. Alonso, C. Benjet, R. Bruffaerts, W. T. Chiu, S. Florescu, G. D. Girolamo, O. Gureje, J. M. Haro, Y. He, C. Hu, E. G. Karam, N. Kawakami, S. Lee, C. Lund, V. Kovess-Masfety, D. Levinson, F. Navarro-Mateu, B. E. Pennell, N. A. Sampson, K. M. Scott, H. Tachimori, M. T. Have, N. C. Viana, D. R. Williams, B. J. Wojtyniak, Z. Zarkov, R. C. Kessler, S. Chatterji and G. Thornicroft, Socio-economic variations in the mental health, *Psychological Medicine*, vol.48, no.9, pp.1560-1571, 2018.

[3] A. Yates, A. Cohan and N. Goharian, Depression and self-harm risk assessment in online forums, *Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp.2968-2978, 2017.

[4] T. Nguyen, D. Phung, B. Dao, S. Venkatesh and M. Berk, Affective and content analysis of online depression communities, *IEEE Transactions on Affective Computing*, vol.5, no.3, pp.217-226, 2014.

[5] R. A. Calvo, D. N. Milne, M. S. Hussain and H. Christensen, Natural language processing in mental health applications using non-clinical texts, *Natural Language Engineering*, vol.23, no.5, pp.649-685, 2017.

[6] K. E. Anderson, Ask me anything: What is Reddit?, *Library Hi Tech News*, vol.32, no.5, pp.8-11, 2015.

[7] J. H. Shen and F. Rudzicz, Detecting anxiety through Reddit, *Proc. of the 4th Workshop on Computational Linguistics and Clinical Psychology – From Linguistic Signal to Clinical Reality*, Vancouver, BC, pp.58-65, 2017.

[8] I. Pirina and Ç. Çöltekin, Identifying depression on Reddit: The effect of training data, *Proc. of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, Brussels, Belgium, pp.9-12, 2018.

[9] M. M. Tadesse, H. Lin, B. Xu and L. Yang, Detection of depression-related posts in Reddit social media forum, *IEEE Access*, vol.7, pp.44883-44893, 2019.

[10] Y. Chen, Y. Guo, H. Jiang, J. Ding and Z. Chen, Self-attention based Darknet named entity recognition with BERT methods, *International Journal of Innovative Computing, Information and Control*, vol.17, no.6, pp.1973-1988, 2021.

[11] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, MN, pp.4171-4186, 2019.

[12] J. A. Naslund, K. A. Aschbrenner and S. J. Bartels, How people with serious mental illness use smartphones, mobile apps, *Psychiatric Rehabilitation Journal*, vol.39, no.4, pp.364-367, 2016.

[13] I. Sekulic, M. Gjurković and J. Šnajder, Not just depressed: Bipolar disorder prediction on Reddit, *Proc. of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Brussels, Belgium, pp.72-78, 2018.

[14] F. M. Shah, F. Ahmed, S. K. S. Joy, S. Ahmed, S. Sadek, R. Shil and M. H. Kabir, Early depression detection from social network, *2020 IEEE Region 10 Symposium (TENSYMP)*, Dhaka, Bangladesh, pp.823-826, 2020.

[15] A. H. Orabi, P. Buddhitha, M. H. Orabi and D. Inkpen, Deep learning for depression detection of Twitter users, *Proc. of the 5th Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, New Orleans, LA, pp.88-97, 2018.

[16] D. William and D. Suhartono, Text-based depression detection on social media posts: A systematic literature review, *Procedia Computer Science*, vol.179, pp.582-589, 2021.

[17] P. He, X. Liu, J. Gao and W. Chen, DeBERTa: Decoding-enhanced BERT with disentangled attention, *arXiv.org*, arXiv: 2006.03654, 2020.