Early detection of depression using BERT and DeBERTa

Sreegeethi Devaguptam, Thanmai Kogatam, Nishka Kotian and Anand Kumar M

Department of Information Technology, National Institute of Technology Surathkal, Karnataka.

Abstract

In today's world, social media usage has become one of the most fundamental human activities. On the report of Oberlo, at present, 3.2 billion people are on social media, which comprises 42% of the World's population. People usually post about their daily life style, special occasions, views about on-going issues and their networks on the social media platforms. People also share things on social media which otherwise would not have shared with other people. Social media helps us to stay connected, keep informed, mobilise on social issues. Due to the surge of suicide attempts, social media can act as a life saver in detecting and tracing users who are on the verge of depression and self-harm. Natural language processing methods with the help of deep learning are aiding in solving language/text related real world problems like sentiment analysis, translation of text into different languages, depression detection. Many transformer based models like BERT (Bidirectional Encoders Representations from Transformers) are put to use to solve NLP problems, which voluntarily learns to attend to different features differently (Weighing). In this paper, a supervised machine learning algorithm with transfer learning approach is used to detect self-harm tendency in the social media users at the earliest.

Keywords

Natural Language Processing, BERT, DeBERTa, transfer learning, text augmentation, social media.

1. Introduction

Major depressive disorder, often known as clinical depression, is a mood condition that affects how you feel and creates a permanent feeling of melancholy and loss of interest. Depression has an impact on how a person feels, thinks, and acts, and can result in a number of emotional and physical issues. According to the World Health Organization's 2020 report, about 264 million people worldwide suffer from depression. It is critical to have it checked; otherwise, there is a risk of depression worsening over time, leading to self-harm or suicide. Major depressive episodes are most common in adolescents aged 12 to 17 years, followed by young adults aged 18 to 25 years, and are rare in people over the age of 50 according to the Substance Abuse and Mental Health Services Association, 2018. Various help lines and mental health awareness initiatives have been established to minimise the suicide rate and to assist persons suffering from depression. Because of modern gadgets, social media networks, and lifestyle, people's behaviours have changed. Millennials prefer to text rather than converse on the phone or face to face. According to a poll of 500 millennials conducted by OpenMarket, "75% of millennials would rather lose the capacity to converse than text". With the growth of social media, there are now a plethora of websites and channels where individuals freely discuss their depression

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

CEUR Workshop Proceedings (CEUR-WS.org)

difficulties and battles. People can share their tales regarding despair, attempted suicide, and other self-harm occurrences in support groups. People prefer to text/write about sadness on social media rather than seek professional treatment, and some even prefer to chat to Bots. We now have some textual data that may be rescued for early diagnosis of depression/self-harming inclinations thanks to social media outlets. Textual data has certain hidden patterns and styles that can be used to identify an author or even gender. Depressed people can be identified using some tools of Natural Language processing. Depressed persons can perhaps be diagnosed using natural language processing techniques before their condition worsens or they self-harm, thanks to improved natural language processing methodologies and tools.

The goal of Early Depression Detection (EDD) is to track a user's messages over time in chronological sequence and determine whether or not the user is depressed early enough to intervene and provide assistance. There haven't been any diversified solutions to the specific challenge to check from other datasets due to the lack of distinct public datasets relating early diagnosis of depression in internet users based on their writings and also feed on their social media accounts. For EDD, researchers have employed Natural Language methods combined with ensemble classifiers, but only a few have attempted to address the problem using cutting-edge deep learning models. Although a machine learning model cannot replace a professional therapist or psychiatrist, not everyone has access to one, and with the proliferation of social media and textual data, a smart model can assist internet users. In this paper, deep learning models are trained for early detection of depression based on texts taken from reddit in chronological order. This paper follows the approach of fine tuning BERT and DeBERTa models along with data augmentation to get a balanced dataset.

2. Related Work

Self-harm detection has got the limelight in deep learning domain as it directly affects the lives of the people. This has attracted many scholars and researchers to work on this problem. There is a huge contribution made in detection of self-harm [8, 9]. Many of the papers have been on different ML (Machine Learning) algorithms in addition with transfer learning models [6, 7].

Yueh-Ming Tai et al. [1] has collected the medical data and eight other factors of Taiwan Soldiers and people admitted in hospitals due to mental disorders like age, depression status, family and community history with depression related problems . They have used Artificial Neural Networks (ANN) algorithm called Radial Basis Function (RBF) models to detect self-harm history and suicide attempt history. Taru Jain et al. [2] has taken tweets from Twitter as their dataset and trained Adversarial Machine Learning algorithm. The paper also talks about the detection of character and word level detection of the self-harm risk. In Adversarial training, they have used augmentation with GANs like SentiGAN.

A. Benton et al. [3] proposes a model which considers the demographic attributes along with mental attributes of a person under study. It has deployed Multi-task Learning (MTL) Framework and logistic regression to detect the onset of different neurological problems in a person. Pratool Bharti et al. [4] uses watch dog model which has three important phases in the detection of self harm risk of a person likely being an accelerometer is given to every studied user to wear it on wrist, developing an to predict whether a person is active/lazy and deploying a machine learning algorithms like random forest classification which can detect the self harm risk in a person.

Parallel efforts have been made to develop strong strategies to improve the robustness of AI systems. Prior research has included word recognition models [13], [14], and denoising auto-encoders [15] to combat character level attacks. Adversarial training has been proven to be effective for word level perturbations [16], and we employ it in our study as well. Other protection strategies include reinforcement learning [17] and detecting adversarial noise before it has an impact on model predictions [18].

3. Methodology

Two transformer based models were trained to detect early traces of depression. From figure 1, it can be seen that the dataset provided is quite imbalanced. It contains a much higher number of writings of users who did not have depression compared to those who did. We applied text augmentation to increase the size of data in the positive class and performed downsampling on the negative class. Figure 3 represents the flowchart of our method (a sample paraphrased sentence is used). A description of the techniques used is given below.

3.1. Data Preprocessing

The first stage is data preprocessing. The writings were cleaned by applying several techniques. Conversion of text to lowercase, removal of urls and html tags, removal of special characters and numbers, stopword removal and substitution of emoticons with their corresponding textual descriptions were included as a part of our preprocessing stage. As most of the writings were within 100 words as seen in Figure 2, longer posts were split into parts to make sure that the length was below 100.

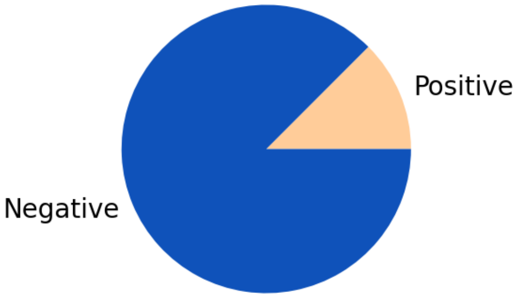

Figure 1: Pie chart of number of users in each class

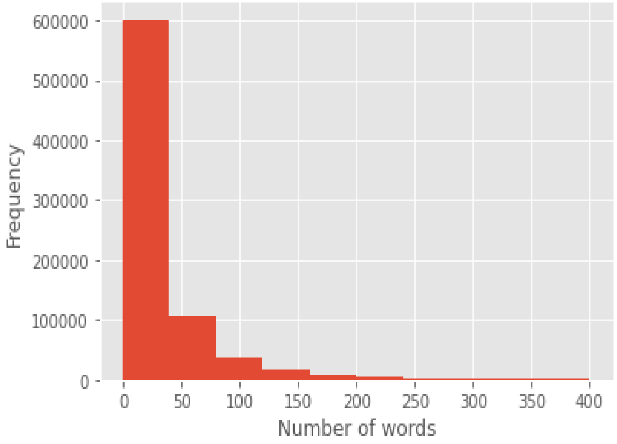

Figure 2: Histogram of post lengths

3.2. Data Augmentation

Data augmentation techniques are used to artificially generate additional data using the available data. It is very widely used in several of the computer vision tasks and can also be used in NLP. Text augmentation is quite challenging as it requires understanding of the context in the sentences. In our method, word level augmentation was applied using the Synonym augmentor from the nlpaug library. The source database used was wordnet and the maximum number of words that would be augmented was set to 20. It simply replaces words with suitable synonyms. This is possibly one of the preferable method in terms of computation cost. Word embedding based augmentation techniques using Glove or even Bert may give better results. Augmentation was done only for the train set.

3.3. Classification

Two transformer based models were trained - namely BERT and DeBERTa. These contain stacked transformer blocks each of which contains a self attention head followed by a fully connected feed forward network. After performing the intial steps, the dataset is divided as 80% data for training the data and 20% for validating the data The input textual data is first converted into the required input data format using a tokenizer. Splitting the tokens is handled by the tokenizer i.e., the sequence is split into tokens available in the tokenizer vocabulary. These tokens can be words or subwords. Additionally, it performs truncation and padding to make sure that the maximum length is 100 and adds special tokens. It returns the input ids and the attention mask in tensor format. Once the input data is prepared, it can be used to train the models.

Transfer learning technique is used wherein the BERT model which was originally pretrained on bookcorpus and wikipedia data on two tasks i.e language modelling and next sentence prediction is fine tuned on our self harm detection dataset to perform sequence classification.

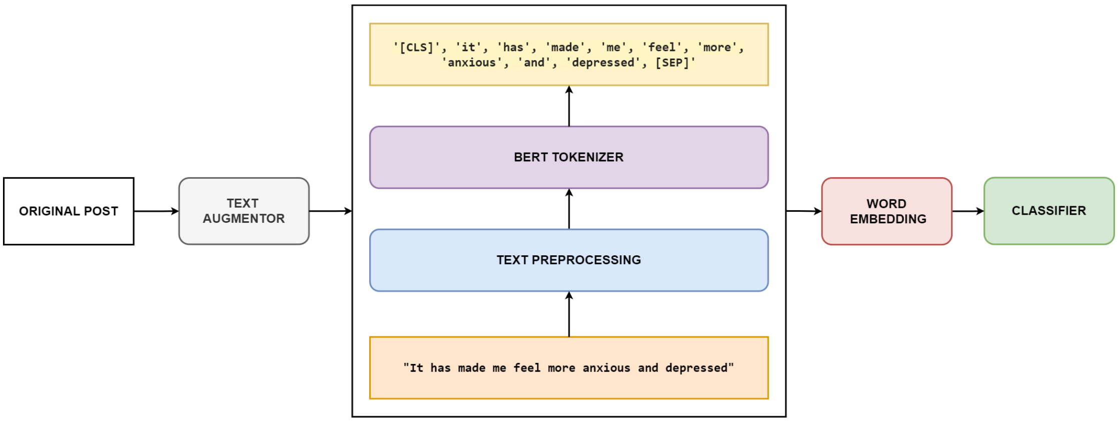
'[CLS]', 'it', 'has', 'made', 'me', 'feel', 'more', 'anxious', 'and', 'depressed', [SEP]'
BERT TOKENIZER
ORIGINAL POST
TEXT AUGMENTOR
WORD EMBEDDING
CLASSIFIER
TEXT PREPROCESSING
"It has made me feel more anxious and depressed"

Figure 3: Flowchart for classification using BERT

3.3.1. BERT

BERT and other Transformer encoder architectures have proven to be quite effective in a range of NLP tasks. They create natural language vector-space representations that can be used in deep learning models. Attention mechanism is used to learn contextual relations between the words. The context of a word is understood from both the left and right parts surrounding it. These models are pre-trained on a huge corpus of text before being fine-tuned for specific tasks. We have used BERT-base cased model for training the model, which has trained weights of the original BERT model. As we are dealing with the binary classification problem i.e., self-harm or not self-harm, we use binary cross-entropy loss function. For optimization, we have used Adam optimizer (learning_rate=2e-5, epsilon=1e-08) which limits the prediction loss and does regularization by weight decay (not utilizing moments).

3.3.2. DeBERTa

DeBERTa improves on the BERT and ROBERTa models. It uses a disentangled attention mechanism where each word is represented in terms of two vectors, one to encode information about the word's position and other about its content. The attention weights are calculated from disentangled matrices on the basis of content and relative arrangement. This is created from the idea that the relative positioning of the words could provide useful information. The dependency between a pair of words could be higher when they occur adjacent to each other. Second, deBERTa uses an enhanced mask decoder which takes into consideration the absolute word positions. Although the relative positional and content are considered by the attention mechanism, it does not take into account the absolute positions which could be pivotal in certain cases. Further, it uses the virtual adversial training regularization technique to improve the performance. It is incorporated right before the softmax layer. This helps to prevent over-fitting and improve generalization. We used SparseCategoricalCrossentropy as the loss function, Adam as the optimizer (learning_rate=2e-05, epsilon=1e-06) and accuracy as the metric.

4. Results

Table 1 contains the performance evaluation results that were obtained. For run 0, BERT model without text augmentation is used, run 1 uses DeBERTa model without text augmentation, run 2 uses BERT model with text augmentation and run 3 uses DeBERTa model with text augmentation. The system fires an alert when the last user post is classified as positive. Comparing the results from these runs it can be found that the highest Recall value was obtained using deBERTa when augmentation was not used while the highest F1 was obtained using deBERTa with augmentation technique. ERDE values were found to be greater in run 2. These values though are not very ideal and there is scope for improvement. There was a considerable time lapse of 01:52:57 while processing the writings which is not optimal. The low number of writings processed (6) may have not been able to give a good idea about the overall model.

Run	P	R	F1	ERDE5	ERDE50	latencyTP	speed	latency-weightedF1
0	0.138	0.796	0.235	0.047	0.039	2.0	0.996	0.234
1	0.135	0.806	0.231	0.047	0.039	2.0	0.996	0.230
2	0.132	0.786	0.225	0.050	0.040	2.0	0.996	0.225
3	0.149	0.724	0.248	0.049	0.039	2.0	0.996	0.247

Table 1
Decision based evaluation result

5. Conclusions and Future work

In this experiment a unique text augmentation method was applied to create additional samples of posts with signs of depression in order to create a balanced dataset. This was done after random downsampling on the negative class. As seen from the results, majority of the performance metrics did not show a great change between the models trained with and without augmentation. The two different transformer models used showed only minor differences in terms of performance. Due to limitations on computing resources the models were trained only for one epoch and also substantial amount of data was removed during the downsampling process applied for the negative class. We have only explored one of the many text augmentation methods. The augmentation process can be refined by creating a pipeline of several different augmentors as well as by the use of more advanced techniques that can take into account the context in the sentences. These aspects can be considered to improve the experiment in the future. Ultimately, the development of an efficient tool that can accurately identify signs of depression from posts can be very beneficial and could be integrated to social media platforms. Further to improve upon the work, an explainable AI model can be developed so that the predicted result can be interpreted and can give an idea about why a post is classified into that particular class.

References

[1] Yueh-Ming Tai; Hung-Wen Chiu: Artificial Neural Network Analysis on Suicide and Self-Harm History of Taiwanese Soldiers, Second International Conference on Innovative Computing, Information and Control (ICICIC 2007).

[2] Taru Jain : Adversarial Machine Learning for Self Harm Disclosure Analysis, 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM).

[3] Adrian Benton, Margaret Mitchell, Dirk Hovy : Multitask learning for mental health conditions with limited social media data, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1 Long Papers, pp. 152-162, Apr. 2017.

[4] Pratool Bharti, Anurag Panwar, Ganesh Gopalakrishna, and Sriram Chellappan : Watch dog : Detecting Self-Harming Activities from Wrist Worn Accelerometers, IEEE Journal of Biomedical and Health Informatics.

[5] Hassan Alhuzali, Tianlin Zhang and Sophia Ananiadou: Predicting Sign of Depression via Using Frozen Pre-trained Models and Random Forest Classifier, CLEF eRisk.

[6] Lu ́ıs Oliveira, Bioinfo@ uavr at erisk 2020: on the use of psycholinguistics features and machine learning for the classification and quantification of mental diseases (2020).

[7] Rodrigo Martínez-Castaño, Amal Htait, Leif Azzopardi, Yashar Moshfeghi : Early risk detection of self-harm and depression severity using bert-based transformers (2020).

[8] Sharath Chandra Guntuku, Daniel Preotiuc-Pietro, Johannes C. Eichstaedt, Lyle H. Ungar : What Twitter profile and posted images reveal about depression and anxiety, Proceedings of the international AAAI conference on web and social media, volume 13, 2019, pp. 236–246.

[9] Mario Ezra Aragón, Adrian Pastor López-Monroy, Luis Carlos González-Gurrola, Manuel Montes-y-Gómez : Detecting depression in social media using fine-grained emotions,Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 1481–1486.

[10] Ana-Maria Bucur , Adrian Cosma and Liviu P. Dinu : Early Risk Detection of Pathological Gambling, Self-Harm and Depression Using BERT, CLEF eRisk 2021.

[11] Elena Campillo-Ageitos , Hermenegildo Fabregat , Lourdes Araujo and Juan Martinez-Romo : NLP-UNED at eRisk 2021: self-harm early risk detection with TF-IDF and linguistic features, CLEF eRisk 2021.

[12] Diana Inkpen, Ruba Skaik, Prasadith Buddhitha, Dimo Angelov and Maxwell Thomas Fredenburgh : uOttawa at eRisk 2021: Automatic Filling of the Beck's Depression Inventory Questionnaire using Deep Learning, CLEF eRisk 2021.

[13] Danish Pruthi, Bhuwan Dhingra, Zachary C. Lipton : Combating adversarial misspellings with robust word recognition, ACL, 2019.

[14] Keisuke Sakaguchi, Kevin Duh, Matt Post and Benjamin Van Durme : Robsut wrod reocginiton via semi-character recurrent neural network, Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence February 4-9 2017, pp. 3281-3287, 2017.

[15] Keita Kurita, Anna Belova, Antonios Anastasopoulos : Towards robust toxic content classification, 2019.

[16] Di Jin, Zhijing Jin, Joey Tianyi Zhou and Peter Szolovits : Is bert really robust*f* a strong

baseline for natural language attack on text classification and entailment, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 8018-8025, 04 2020.

[17] Jingjing Xu, Liang Zhao, Hanqi Yan, Qi Zeng, Yun Liang and Xu Sun, "LexicalAT: Lexical-based adversarial reinforcement training for robust sentiment classification", Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Nat-ural Language Processing (EMNLP-IJCNLP), pp. 5518-5527, Nov. 2019.

[18] Yichao Zhou, Jyun-Yu Jiang, Kai-Wei Chang, Wei Wang : Learning to discriminate perturbations for blocking adversarial attacks in text classification, EMNLP/IJCNLP, 2019.

[19] Pengcheng He, Xiaodong Liu, Jianfeng Gao, Weizhu Chen : DeBERTa: Decoding-enhanced BERT with Disentangled Attention ,ICLR, 2021.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, Attention Is All You Need, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL 2019.

[22] Javier Parapar, Patricia Martín-Rodilla, David E. Losada, Fabio Crestani: Overview of eRisk 2022: Early Risk Prediction on the Internet. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 13th International Conference of the CLEF Association, CLEF 2022. Springer International Publishing, Bologna, Italy.