KADO@LT-EDI-ACL2022: BERT-based Ensembles for Detecting Signs of Depression from Social Media Text

Morteza Janatdoust
Amirkabir University of Technology
jntdst@aut.ac.ir

Fatemeh Ehsani-Besheli
K. N. Toosi University of Technology
ehhsani@aut.ac.ir

Hossein Zeinali
Amirkabir University of Technology
hzeinali@aut.ac.ir

Abstract

Depression is a common and serious mental illness that early detection can improve the patient’s symptoms and make depression easier to treat. This paper mainly introduces the relevant content of the task “Detecting Signs of Depression from Social Media Text at DepSign-LT-EDI@ACL-2022”. The goal of DepSign is to classify the signs of depression into three labels namely “not depressed”, “moderately depressed”, and “severely depressed” based on social media’s posts. In this paper, we propose a predictive ensemble model that utilizes the fine-tuned contextualized word embedding, ALBERT, DistilBERT, RoBERTa, and BERT base model. We show that our model outperforms the baseline models in all considered metrics and achieves an F1 score of 54% and accuracy of 61%, ranking 5th on the leaderboard for the DepSign task.

Keywords. sentiment analysis, depression detection, ensemble model, BERT, social media text

1 Introduction

In our current society, depression is a common but serious mental disorder that involves sadness and lack of interest in all day-to-day activities (GHD; Evans-Lacko et al., 2018). Depression can negatively affect different aspects of a person’s life and can cause a person to suffer severely and function poorly at work, in the family, or in society in general and at its worst, depression can lead to suicide. Based on the data provided by World Health Organization, Over 700,000 people die due to suicide every year (WHO). Therefore, early diagnosis of this problem is very important and is a challenge for individual and public health (Losada et al., 2017).

Because of the complex nature of any mental disorder, it is very difficult to diagnose a patient’s mental illness by traditional approaches. However, due to the integration of social media into people’s daily lives, evidence has been presented to diagnose depressive symptoms using data provided by users.

The study of social media, especially in the field of public health, is rapidly growing. On social media platforms such as Facebook, Twitter, Instagram, and others, people can freely interact with each other and share their thoughts, feelings, ideas, emotions, activities, etc and express themselves through the content they post on these platforms. This leads to a large amount of data that contains valuable information about people’s interests, moods, and behaviors. Hence many researchers claim that social media analysis is a very helpful source in various contexts especially in mental health understanding (Martínez-Castaño et al., 2020).

2 Related Work

There have been many studies on the prediction of social media mental disorders in which the data were collected directly from user surveys using some well-known questionnaires or from public posts using keywords, related phrases, or regular expression (Safa et al., 2021). Several approaches to study mental health have been proposed through the analysis of user behavior on social media. Mental health has been studied on different social media platforms such as Twitter, Instagram, Flickr, and Facebook. In (Orabi et al., 2018), using a deep neural network, an analysis was performed to diagnose depression on the Twitter database. (De Choudhury et al.), has also analyzed Twitter social media text for public health prediction.

Binary and ternary classifications are two types of classification problems here. In the first one, sentiments are classified into two polarities or classes: Positive and Negative (Tanna et al., 2020), and in the ternary classification, the sentiments are classified into three classes as Positive, Negative and Neutral (Arora and Arora, 2019; Chen et al., 2018) which in this case, more classification error is expected than binary classification.
For more accurate classification, the data can be classified into several subclasses. In (Al Asad et al., 2019) for example, having a level of depression from 1-55% is considered as non-depressed and above level of 55% is considered as depressed. The defined subclasses were normal, mild depression, borderline depression, moderate depression, and severe depression.

In this article, we specifically focus our efforts on this kind of classification task. Our goal is to distinguish between the normal users, users with mild depression, and those with severe depression.

| Label          | Train | Dev |
|----------------|-------|-----|
| Not depressed  | 1,971 | 1,830 |
| Moderate       | 6,019 | 2,306 |
| Severe         | 901   | 360  |
| Total instances| 8,891 | 4,496 |

Table 1: Train and Validation data-sets description.

2.1 Data

The data-test provided by the organizer (Durairaj et al.), contains social media comments in English. It comprises training, development and test set, in which 8,891 are assigned for training, 4,496 for development, and 3,245 for testing. The data set contains three tags as follow:

- not depressed: This tag indicates that sentence shows the absence of depression,
- moderately depressed: This tag indicates depressive symptoms,
- severely depressed: This tag indicates severe states of depressed mood.

Figure 1 illustrates how the different classes are represented in the data sets, which shows that the distribution of examples in the classes are imbalanced for both train and development data-set. The details of the data-set and three example sentences are shown in Table 1 and Table 2 respectively.

3 Transfer Learning

Typically, models are trained from scratch with random initialization of network parameters. But in another approach, the model is first pre-trained for a general task and then tuned to a specific task, which allows the model to be trained faster with less training data. Originally, transfer learning is known for fine-tuning the deep learning models taught on the ImageNet data-set (Deng et al., 2009). Recently, several techniques and architectures of transfer learning have been emerged, which has significantly improved most NLP tasks. Transfer learning can be used in applications where there is not sufficient training data for that task. The first phase of the transfer learning strategy is generally referred to as semi-supervised training in which the network is first trained as a language model on a comprehensive and large data set and then followed by supervised training that is trained by the desired labeled training data set.

3.1 BERT

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) is a deep transformer model designed to learn deep bidirectional representations of natural language from a huge unsupervised text corpus. In terms of size, there are three BERT models. The base model consists of 12 transformer blocks, 768 hidden blocks, 12 self-attention heads, and has 110M trainable parameters.

BERT uses two tasks called Masked Language Model(MLM) and Next Sentence Prediction(NSP) to train the model. In the MLM task, before feeding word sequences into BERT, 15% of tokens are covered by [MASK] token and the model tries to predict the original value of the covered token based on non-masked words in the input sequence. In the NSP task, the BERT model takes a pair of sentences as input and, by understanding the relationship between two sentences, predicts if the second sentence in the pair is the subsequent sen-
My life gets worse every year: That’s what it feels like anyway....
Words can’t describe how bad I feel right now: I just want to fall asleep forever.
Is anybody else hoping the Coronavirus shuts everybody down?

| Model        | Accuracy | Recall | Precision | Weighted F1-score | Macro F1-score |
|--------------|----------|--------|-----------|--------------------|----------------|
| BERT         | 0.55     | 0.52   | 0.48      | 0.56               | 0.50           |
| ALBERT       | 0.56     | 0.51   | 0.50      | 0.57               | 0.51           |
| DistilBERT   | 0.52     | 0.50   | 0.48      | 0.59               | 0.49           |
| RoBERTa      | 0.57     | 0.53   | 0.51      | 0.60               | 0.52           |
| Ensemble Model | 0.61   | 0.57   | 0.52      | 0.62               | 0.54           |

Table 3: Label-averaged values for each metric for BERT-based model, ALBERT, RoBERTa, DistilBERT, and the proposed ensemble model.
strengths and weaknesses as different classification methods. As a result, ensembling them can improve the result.

Each constituent model is trained on a pretty same training data and the same loss function is used for parameter estimation of each model. Our experiments showed that each of these models makes different errors. So, for this problem, we have used the majority voting mechanism to make the final prediction to use the strength of each model (see in Figure 2).

![Figure 2: Architecture of the proposed ensemble model.](image)

Before fine-tuning each of the pretrained models, the proper number of epochs must be known. Using a validation data-set that is held back from training, we identify overfitting by looking at validation metrics like loss and accuracy and define correct number of epochs for training each model. Whenever the loss value in the validation set increases, it means that network training should be stopped by this number of epochs. From then on, given the data-set for this task, we train the model and tune it to the predefined number of epochs to perform well on unseen data points.

### 4.1 Results

After training the ensemble model, it was evaluated with the test data-set. Depending on the number of models in the voting-based ensemble model, the same number of test data answers are obtained. Then, the unlabeled test set can be classified by the majority voting ensemble learning method. The accuracy results obtained on the evaluation data-set for all models are shown in Table 1. The results show that the ensemble-based model utilizing contextual embeddings outperforms other single-model classifiers in all considered metrics and achieves an F1 score of 54% and accuracy of 61%.

### 4.2 Conclusion

This paper presents a BERT-based ensemble model to predict depression levels based on the given labels: not depressed, moderately depressed, and severely depressed. The proposed ensemble model achieved competitive results for the label prediction on the DepSign task and ranked 5th among more than 30 submissions. By considering the achieved improvement, future works could be examining other language models, other ensemble strategies, and use other inputs such as related dictionaries, NLP tools, and etc.

### References

Institute of Health Metrics and Evaluation (IHME). global health data. [http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/d780dffe8631b8e6a1f1e48959e88be](http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/d780dffe8631b8e6a1f1e48959e88be).

World Health Organization (WHO), geneva, switzerland. [https://www.who.int/news-room/factsheets/detail/suicide](https://www.who.int/news-room/factsheets/detail/suicide). Accessed: 2021-06-17.

Nafiz Al Asad, Md Appel Mahmud Pranto, Sadia Afreen, and Md Maynul Islam. 2019. Depression detection by analyzing social media posts of user. In 2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICS CON), pages 13–17. IEEE.

Priyanka Arora and Parul Arora. 2019. Mining twitter data for depression detection. In 2019 International Conference on Signal Processing and Communication (ICSC), pages 186–189. IEEE.

Bohang Chen, Qiongxia Huang, Yiping Chen, Li Cheng, and Riqing Chen. 2018. Deep neural networks for multi-class sentiment classification. In 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), pages 854–859. IEEE.

M De Choudhury, M Gamon, S Counts, and E Horvitz. Predicting depression via social media. 2013 jul presented at: Proceedings of the seventh international aaai conference on weblogs and social media; july; 2013. Cambridge, Massachusetts, USA, pages 128–137.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee.

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.
Findings of the shared task on Detecting Signs of Depression from Social Media.

S. Evans-Lacko, S. Aguilar-Gaxiola, A. Al-Hamzawi, J. Alonso, C. Benjet, R. Bruffaerts, W. T. Chiu, and S. Florescu. 2018. Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: Results from the who world mental health (wmh) surveys. *Psychological Medicine*, 48:1560–1571.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

David E Losada, Fabio Crestani, and Javier Parapar. 2017. erisk 2017: Clef lab on early risk prediction on the internet: experimental foundations. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 346–360. Springer.

Rodrigo Martínez-Castaño, Juan C Pichel, and David E Losada. 2020. A big data platform for real time analysis of signs of depression in social media. *International Journal of Environmental Research and Public Health*, 17(13):4752.

Ahmed Hussein Orabi, Prasadith Buddhitha, Mahmoud Hussein Orabi, and Diana Inkpen. 2018. Deep learning for depression detection of twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97.

Ramin Safa, Peyman Bayat, and Leila Moghtader. 2021. Automatic detection of depression symptoms in twitter using multimodal analysis. *The Journal of Supercomputing*, pages 1–36.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Dilesh Tanna, Manasi Dudhane, Amrut Sardar, Kiran Deshpande, and Neha Deshmukh. 2020. Sentiment analysis on social media for emotion classification. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 911–915. IEEE.