iicteam@LT-EDI: Leveraging pre-trained Transformers for Fine-Grained Depression Level Detection in Social Media

Vajratiya Vajrobol, Karanpreet Singh, Nitisha Aggarwal
Institute of Informatics and Communication, University of Delhi, India.
tiya101@south.du.ac.in,
karanpreet.singh@iic.ac.in, nitisha@south.du.ac.in

Abstract

Depression is a significant mental illness characterized by feelings of sadness and a lack of interest in daily activities. Early detection of depression is crucial to prevent severe consequences, making it essential to observe and treat the condition at its onset. At ACL-2022, the DepSign-LT-EDI project aimed to identify signs of depression in individuals based on their social media posts, where people often share their emotions and feelings. Using social media postings in English, the system categorized depression signs into three labels: "not depressed," "moderately depressed," and "severely depressed." To achieve this, our team has applied MentalRoBERTa, a model trained on big data of mental health. The test results indicated a macro F1-score of 0.439, ranking the fourth in the shared task.

1 Introduction

Depression is a significant mental health condition that affects individuals worldwide. It is characterized by persistent feelings of sadness, lack of interest in daily activities, and a range of emotional and physical symptoms. The World Health Organization (WHO) estimates that more than 300 million people of all ages suffer from depression, making it a significant public health concern (World Health Organization, 2017) (Organization, 2017). Early detection and intervention play a crucial role in effectively addressing depression and reducing its impact on individuals' lives (Beames et al., 2021).

Recognizing the importance of early detection, researchers have turned to artificial intelligence (AI) techniques to develop automated systems for the detection of depression. Social media platforms have emerged as a valuable source of data for understanding individuals' mental health, as people often express their thoughts, emotions, and experiences in their online posts. By leveraging the vast amount of user-generated content on social media, researchers aim to detect signs of depression in a timely manner and provide appropriate support (D'Alfonso, 2020). In this research, we focus on leveraging pre-trained learning models, specifically MentalRoBERTa transformers, for fine-grained depression detection in social media. Pre-trained transformers have shown remarkable success in various natural language processing tasks, and we seek to harness their capabilities to accurately identify and classify depression-related patterns in social media text (Vajrobol et al., 2022).

One of the challenges in depression detection is the presence of imbalanced datasets, where instances of non-depressive posts and moderate level of depression are significantly outnumbered by severe depression-related posts. To address this issue, we employ text augmentation techniques to artificially increase the number of depressive instances in the dataset, thereby enhancing the model's ability to learn from the imbalanced data distribution.

The primary contribution of our research lies in developing a robust model for fine-grained depression detection by leveraging pre-trained MentalRoBERTa transformers and employing text augmentation techniques to handle imbalanced datasets. By detecting depression at a fine-grained level, we aim to provide more nuanced insights into individuals' mental health states and facilitate targeted interventions and support.

In the following sections, we will describe the related work in the field of depression detection from social media and discuss the methodolo-

LT-EDI 2023 – Third Workshop on Language Technology for Equality, Diversity and Inclusion,
pages 83–88, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-084-7_012

gies and experiments conducted in our research. The results obtained will demonstrate the effectiveness and contributions of our proposed approach for leveraging pre-trained learning in fine-grained depression detection. Ultimately the conclusion and future works will be drawn.

2 Literature surveys

Depression is a significant mental health issue that affects millions of people worldwide. Early detection and intervention are crucial for providing timely support to individuals suffering from depression. With the rise of social media platforms, researchers have begun exploring the potential of leveraging pre-trained learning models for fine-grained depression detection in social media posts. In a recent study, (Lam et al., 2019) employed multi-modal data and proposed a novel method that combines topic modeling using transformers and a deep 1D convolutional neural network (CNN) for acoustic feature modeling. The results demonstrated that the deep 1D CNN and transformer models achieved state-of-the-art performance for audio and text modalities, respectively. Furthermore, the multi-modal results are comparable to the state-of-the-art depression detection systems.

Furthermore, (Martnez-Castano et al., 2020) investigated early detection of signs of self-harm and measuring the severity of the signs of Depression. Their approach focused on utilizing BERT-based classifiers trained specifically for each task. The results demonstrated that this approach yielded excellent performance across various measures. The study suggested that trained BERT-based classifiers can accurately identify social media users at risk of self-harming, with a precision rate of up to 91.3 %. Recent study also performed depression detection in low-resource language like Thai, and found out XLM-RoBERTa based on Transformers has performed the best with 79.12% accuracy (Vajrobol et al., 2022).

A study by (Meng et al., 2021) focused on leveraging the application of temporal deep learning models with a transformer architecture to predict future diagnosis of depression using electronic health record (EHR) data. The model demonstrated improved precision-recall area under the curve (PRAUC) for depression prediction, achieving a PRAUC increase from 0.70 to 0.76 compared to the best baseline model.

(S et al., 2022) investigated the detection of signs of depression in social media Text using transformer models like DistilBERT, RoBERTa, and ALBERT. The prediction process involved assigning three labels to the data: Moderate, Severe, and Not Depressed. The evaluation of their models revealed Macro F1 scores of 0.337, 0.457, and 0.387 for DistilBERT, RoBERTa, and ALBERT, respectively.

One notable study by (Zhang et al., 2022) focused on utilizing a hybrid deep learning model called RoBERTa-BiLSTM to extract features from sequences of depression text. The model combines the strengths of sequence models and Transformer models while mitigating the limitations of sequence models. By utilizing the Robustly optimized BERT approach, the model maps words into a meaningful word embedding space and effectively captures long-distance contextual semantics using the Bidirectional Long Short-Term Memory model. Experimental results demonstrated that this model holds promise for improving the accuracy and robustness of depression detection, aiding in the timely identification and treatment of individuals experiencing depression.

Another relevant study by (Vetulani et al., 2023) demonstrated that transformer ensembles outperformed individual transformer-based classifiers in detecting depression. This finding underscores the significance of leveraging ensemble models to improve the accuracy and robustness of depression detection from social media posts. By harnessing the power of transfer learning, we can effectively apply knowledge gained from one dataset to enhance the performance on a different dataset, expanding the applicability of our models.

These previous studies collectively illustrated the effectiveness and versatility of leveraging pre-trained learning for fine-grained depression detection in social media. By fine-tuning pre-trained transformers, researchers have been able to capture intricate linguistic patterns and emotional expressions indicative of depression. This approach holds immense promise for developing accurate and scalable tools for early detection and intervention in individuals at risk of depression.

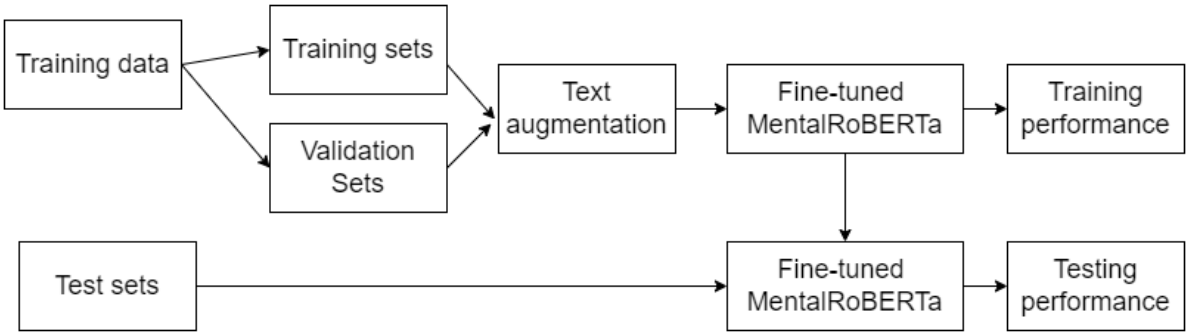

Figure 1: The process of detection level of depression

3 Methods

The Training data has been provided by competition organizers, further divided into training sets and validation sets 80:20 as demonstrated in Figure 1. After splitting dataset, text augmentation techniques have been applied because the shared training data is imbalanced. For the classification, MentalRoBERTa has been trained and fine-tuned on training sets. Furthermore, this fine-tuned MentalRoBERTa has been utilized to assess testing performance with test sets.

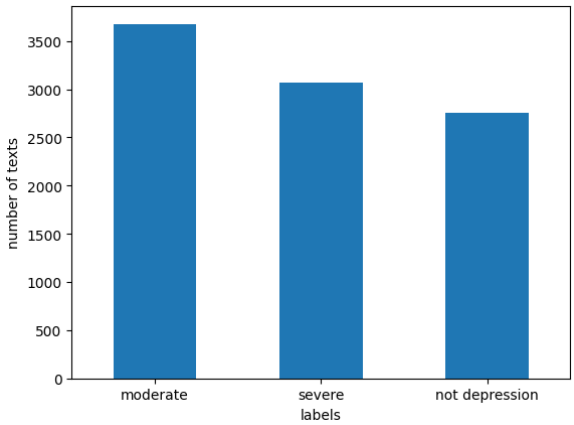

Figure 2: The distribution of the level of depression dataset after augmentation.

3.1 Dataset and data pre-processing

The original training dataset in the shared task (Sampath et al., 2023) has been included in 7,201 records and divided into three labels, such as moderate depression with 3,678 rows, not depression with 2,755 rows, and severe depression with 768 rows (Losada et al., 2017; DravidianLangTech, 2023). The data is hugely imbalanced. Therefore, we have applied two text augmentation techniques like synonyms (replacing words with similar meanings) and random swap (rearranging word order) enhance data variety, aiding machine learning models to better understand language and generalize effectively. Finally, the whole dataset has been added up to 9,505 rows, which include 3,678 records with moderate depression label, 2,755 records with non-depression label, and 3,072 records with severe depression label as it can be seen in Figure 2 . And an example of a dataset has been shown in Table 1. The test dataset included 499 records.

3.2 MentalRoBERTa

MentalRoBERTa is a pre-trained language model specifically developed for mental health-related natural language processing tasks. The training corpus for MentalRoBERTa was collected from Reddit, an network of communities where users engage in discussions on various topics of interest. For the purpose of focusing on the mental health domain, several relevant subreddits were selected to crawl users' posts. During the data collection process, user profiles were not collected, even though they are publicly available on Reddit. The aim was to ensure user privacy and adhere to ethical considerations. The selected mental health-related subreddits include "r/depression," "r/SuicideWatch," "r/Anxiety," "r/offmychest," "r/bipolar," "r/mental illness," and "r/mentalhealth." These subreddits provide a diverse range of discussions related to mental health, covering topics such as depression, anxiety, bipolar disorder, and general

Text	Label
What to do these days?: I've struggled with Depression and anxiety for all my life now and every time I'd feel alone or down there was always something I could do. Usually when I'd want to feel better I'd go to a cafe and read or just chill to calm down or go look for cool things in stores. I could even go to the swimming pool or gym.	SEVERE
2019 was my worst year with 2 depression crises. I'm happy it ended but so afraid of what 2020 will bring. : This year was rough. It started on the NYE with my puppy almost dying from the fireworks. He literally shat all over myself. In may I had my first terrible depression and anxiety crisis and had to be away from my internship for 2 weeks. Then in June I went through the first real loss of my life. My dear uncle died from a heart attack all of sudden.	MODERATE
Have a happy near year.... : I'm spending this new year alone and in bed. I hope you are not doing the same. I hope you can have fun today if you're reading this. Next year is gonna be lit, dont give up on yourself. You're all you got in this life.	NOT DEPRESSION

Table 1: The example of the training dataset.

mental health issues. The resulting training corpus for MentalRoBERTa consists of a total of 13,671,785 sentences. This corpus encompasses a broad range of textual expressions, including personal narratives, experiences, questions, and support-seeking posts related to mental health. By training MentalRoBERTa on this extensive dataset, the model can effectively learn and capture the language patterns, context, and semantics specific to mental health discussions on social media. MentalRoBERTa, being pre-trained on this mental health-specific corpus, enables researchers and practitioners to leverage its knowledge and capabilities in various natural language processing tasks related to mental health. This includes sentiment analysis, mental health classification, identification of specific mental health conditions, and other text-based analyses in the field of mental healthcare (Ji et al., 2022).

4 Results and Discussion

The training loss is presented at different steps of the training process in Figure 3. At step 500, the training loss is 0.8063, indicating that the model's predictions have a relatively higher deviation from the actual target values. As the training progresses, the training loss decreases, indicating that the model is improving its performance and fitting the training data better. At step 2500, the training loss is 0.5932, showing a further reduction in the loss value. Monitoring the training loss helps assess the convergence and performance of the model during training. Lower training loss values generally indicate a better fit to the training data. However, it is important to note that the training loss alone may not fully reflect the model's performance on unseen data or in real-world scenarios. Evaluation on separate validation or test datasets is necessary to assess the model's generalization ability and overall performance. The training performance results show that the model generated accuracy 69.96 %, F1-score

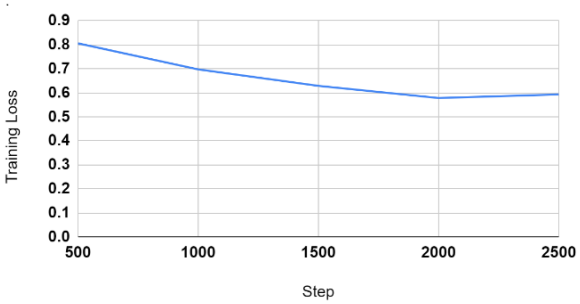

Figure 3: Training loss and steps in the training model process.

69.94 %. Then we evaluated the model and using the 499 samples of the test set, the results shows that a macro F1-score of 0.439, as the fourth-ranked participant in the shared task.

5 Conclusion

The DepSign-LT-EDI project focused on the detection of depression signs in individuals based on their social media postings. By utilizing the MentalRoBERTa model trained on mental health data, the model classified the signs of depression into three labels: "not depressed," "moderately depressed," and "severely depressed." The result obtained from the evaluation of the system showcased a macro F1-score of 0.439, positioning the system as the fourth-ranked participant.

Another area for future investigation involves incorporating multimodal analysis. By integrating textual analysis with other modalities such as images, videos, and audio, a more holistic understanding of individuals' mental health states can be achieved. This multimodal approach has the potential to improve the accuracy and reliability of depression detection systems.

Furthermore, there is room for refining the classification system to capture finer levels of depression severity. Developing models that can distinguish between different levels of depression, ranging from mild to moderate and severe, would enable a more nuanced understanding of individuals' mental well-being. This, in turn, could facilitate more targeted interventions and personalized support. It is also crucial to consider ethical considerations in future research endeavors. Addressing privacy concerns, obtaining informed consent, and mitigating potential biases in depression detection from social media are essential. Striving for transparency and interpretability in the developed models while ensuring data protection and respecting individuals' autonomy is of utmost importance. By exploring these future directions, the field of depression detection from social media can continue to advance. This progress would lead to the development of more accurate and effective tools for early intervention, support, and treatment for individuals experiencing depression.

Acknowledgments

The authors would like to thank Project Samarth, an initiative of the Ministry of Education(MoE), Government of India, at the University of Delhi South Campus (UDSC), for their support

References

Joanne R. Beames, Katarina Kikas, and Aliza Werner-Seidler. 2021. Prevention and early intervention of depression in young people: an integrated narrative review of affective awareness and ecological momentary assessment. *BMC Psychology*, 9:113.

DravidianLangTech. 2023. Detecting signs of depression from social media text - lt-edi@ranlp 2023.

Simon D'Alfonso. 2020. Ai in mental health. *Current Opinion in Psychology*, 36:112–117.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. Mentalbert: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190. European Language Resources Association.

Genevieve Lam, Huang Dongyan, and Weisi Lin. 2019. Context-aware deep learning for multimodal depression detection. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3946–3950. IEEE.

David E Losada, Fabio Crestani, and Javier Parapar. 2017. erisk 2017: Clef lab on early risk prediction on the internet: experimental foundations. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings 8*, pages 346–360.

Rodrigo Martnez-Castano, Amal Htait, Leif Azzopardi, and Yashar Moshfeghi. 2020. Early risk detection of self-harm and depression severity using bert-based transformers. *Working Notes of CLEF*, page 16.

Yiwen Meng, William Speier, Michael K Ong, and Corey W Arnold. 2021. Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. *IEEE journal of biomedical and health informatics*, 25:3121–3129.

World Health Organization. 2017. "depression: let's talk" says who, as depression tops list of causes of ill health.

Sivamanikandan S, Santhosh V, Sanjaykumar N, Jerin Mahibha C, and Thenmozhi Durairaj. 2022. scubemsec@lt-edi-acl2022: Detection of depression using transformer models. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 212–217. Association for Computational Linguistics.

Kayalvizhi Sampath, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Jerin Mahibha C, Kogilavani Shanmugavadivel, and Pratik Anil Rahood. 2023. Overview of the second shared task on detecting signs of depression from social media text. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Vajratiya Vajrobol, Unmesh Shukla, Amit Pundir, Sanjeev Singh, and Geetika Jain Saxena. 2022. Depression detection in thai language posts based on attentive network models. *CEUR Workshop Proceedings*.

Zygmunt Vetulani, , and Patrick Paroubek and, editors. 2023. *Human Language Technologies as a Challenge for Computer Science and Linguistics – 2023*. Adam Mickiewicz University Press.

Yazhou Zhang, Yu He, Lu Rong, and Yijie Ding. 2022. A hybrid model for depression detection with transformer and bi-directional long short-term memory. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2727–2734. IEEE.