IIITDWD@LT-EDI-2023 Unveiling Depression: Using pre-trained language models for harnessing domain-specific features and context information

Shankar Biradar[1] and **Sunil Saumya**[1] and **Sanjana Kavatagi**[2]
[1]Department of Computer Science and Engineering,
Indian Institute of Information Technology, Dharwad, Karnatka, India
[2]Department of Computer Science and Engineering, VTU, Belagavi
(shankar,sunil.saumya)@iiitdwd.ac.in
kawatagi.sanjana@gmail.com

Abstract

Depression is a global health crisis affecting millions. Workplace stress and unhealthy habits have risen, leading to more people with depressive symptoms. Early detection and prediction of depression are essential for timely intervention and support. Unfortunately, social stigma prevents many from seeking help, making early detection difficult. Therefore, alternative strategies for depression prediction, such as analysing social media posts, are being explored. LT-EDI@RANLP held a shared task to promote research in this field. Our team participated in the shared task and secured 21st rank with a macro F1 score of 0.36. This article summarises the model used in the shared task.

1 Introduction

Depression is a common mental health illness affecting millions worldwide, causing significant suffering and even terrible outcomes such as suicide. Despite its frequency and negative impact, depression frequently remains unrecognized and untreated, particularly among young adults. It is essential to understand the wide-ranging harmful effects of this invisible killer to address the worldwide mental health crisis. Over 280 million people worldwide suffer from depression, and the number is rising, according to World Health Organisation (WHO) [1]. The effects of depression are disastrous for both mental and physical health. It limits a person's ability to succeed in life, including work, relationships, and personal fulfilment. Additionally, depression ranks as the second leading cause of teenage mortality, highlighting the urgent need for improved detection and treatment strategies[2].

Traditional diagnostic procedures, which rely on patient self-reports, remarks from family or friends, and mental state examinations, frequently encounter major problems. There are many people who are depressed who do not receive the appropriate treatment because of underdiagnosis, undertreatment, cultural stigma, and inaccurate assessments. The rise of social media platforms like Facebook, Twitter, and WhatsApp in recent years has provided new avenues for understanding mental health conditions like depression (Coppersmith et al., 2014; Lin et al., 2016; Biradar et al., 2022). More and more people are using these platforms to express their thoughts, feelings, and everyday experiences, which tells us a lot about their mental health. By leveraging user behaviours, language patterns, and social connections, researchers are exploring the potential of social media as a tool for diagnosing and forecasting depression.

The COVID-19 epidemic has underlined the necessity of technology-based interventions in mental healthcare. With the adoption of social distancing measures and lockdowns, people have resorted to social media for communication, self-expression, and support. The pandemic's impact on mental health, limited resources, and overworked healthcare systems have highlighted the need for creative and scalable methods for depression detection and treatment. Overall, depression remains a substantial global concern, demanding novel techniques for identification and treatment. The combination of social media and behavioural factors offers a promising path for forecasting depression levels. The advancement of technology, including artificial intelligence and machine learning techniques, presents an intriguing potential for leveraging the massive volumes of data available on social media networks. We can use these technologies to create strong, personalized systems that help healthcare professionals, researchers, and individuals identify and treat depression more effectively (Akbari et al., 2016; Kayalvizhi et al., 2022; Chakravarthi et al.,

[1] https://www.who.int/news-room/fact-sheets/detail/depression

[2] https://www.who.int/news-room/fact-sheets/detail/depression

LT-EDI 2023 – Third Workshop on Language Technology for Equality, Diversity and Inclusion, pages 117–123, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-084-7_017

2022).

Several methods for handling social media data to determine users' depression conditions have been presented. However, most of these approaches have relied on handcrafted features with shallow machine learning-based models (Tadesse et al., 2019; Guntuku et al., 2019; Biradar et al., 2021). These approaches often require domain expertise to identify features, resulting in biased feature values. Furthermore, the handcrafted feature extraction method is laborious and time-consuming, resulting in a longer training time. In addition, many of these models struggle to generalize successfully with new information. Researchers have recently tried to alleviate these constraints by employing pre-trained transformer models (Poerner et al., 2020; Kassner and Schütze, 2020; Puranik et al., 2021). However, to the best of our knowledge, none of these models have successfully linked domain knowledge with linguistic patterns. To overcome this gap, our proposed work performed experiments with PubMed BERT (Gu et al., 2020) trained on clinical data, to harness domain knowledge. These studies were carried out as part of the LT-EDI@RANLP joint task on Detecting Signs of Depression from social media Text. Notably, our proposed model finished in the 21st position among the participating teams.

The remaining part of paper is arranged as follows: Section 2 addresses the recent literature. Further model building details are discussed in section 3. Finally, section 4 provides insights into the model results. Furthermore, its implications on society and future research directions are provided in the last section.

2 Background study

Depression detection addresses the interdisciplinary topic of clinical psychology and social media data mining. Several studies have been proposed to analyze social media users' behaviour through their content. The results from these studies conclude that individuals with depression tend to use more negative verbal content when interacting with friends or posting on social media. Most of these models are conducted using either machine learning-based or deep learning-based methods. This section will provide insights into some selected works from the past.

2.1 Machine learning-based models

(Tadesse et al., 2019; Shankar Biradar and Chauhan, 2021) Conducted an experiment involving n-gram features representation, such as tf-idf, linguistic features, and LDA topics. The study utilized Logistic Regression (LR), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Random Forest (RF), and AdaBoost to train the models. The experimental results indicated that SVM achieved an accuracy of 81%. However, the highest performance was achieved using the Multi-Layer Perceptron, with an accuracy of 91% for depression and non-depression classification. Several studies have also focused on syntactic and semantic features for detecting depressive comments from social media data. Liu and Shu extracted syntactic and semantic-based features to train several supervised learning models, such as Naive Bayes (NB), K-Nearest Neighbors (KNN), Logistic Regression, and Support Vector Machine, as base learners. Later, a simple logistic regression model was used to stack the outcomes of the base learners (Liu and Shi, 2022).

In a study (Tsugawa et al., 2015), semantic features were found to be integral components of depression prediction models. The researchers utilized various semantic features and word-level attributes to gauge the level of depression among Twitter users. These features included word frequency and the ratio of positive to negative words. By employing SVM classifiers, the study demonstrated that semantic features could effectively address depressive comments on social media. The findings indicate that semantic features hold promise in identifying and handling instances of depression on online platforms. In a related study (Pirina and Çöltekin, 2018; Shankar Biradar and Chauhan, 2021), the authors trained an SVM classifier using various word-level features to identify the severity of depressive comments. The study utilized features such as word n-grams and tf-idf for training LR, RF, and SVM classifiers. The study concluded that the combination of word-level features with the SVM model yielded superior results in predicting depression levels from social media comments. (Chen et al., 2018) developed a binary classifier for depression detection and achieved great accuracy by utilizing Random Forests and Support Vector Models with Radial Basis Function.

2.2 Deep Learning-based models

Recent studies have found that deep learning-based methods can significantly enhance model performance. In addition to this, DL models also reduce the computational overhead of ML-based models during the feature extraction stage. Traditional feature extraction in ML models requires domain expertise and is time-consuming, often leading to biased features. To address these issues, several studies have proposed the use of deep learning-based models.

For instance, (Wani et al., 2022) extracted word-level embeddings using a pre-trained word2vec neural network model. These embeddings were then passed to Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM) for classification. The results of the study concluded that using RNN-based methods improves model performance. In another study, authors attempted to build their own corpus containing binary depressive and non-depressive comments (Kim et al., 2020). They employed word2vec embeddings combined with a CNN model for depression detection. Some researchers also explored the construction of hybrid models by combining CNN and LSTM networks (Kour and Gupta, 2022; Biradar and Saumya, 2022). These hybrid networks successfully capture spatial features (CNN) and temporal features (LSTM) to address depression levels in long text. The study concluded that using hybrid networks improves model performance. Lastly, given the prevalence of COVID-19-related depressive comments on social media, researchers (Zogan et al., 2023) developed a corpus specifically related to COVID-19 depressive comments. They also presented a novel hierarchical CNN network for binary classification. These advancements in deep learning-based approaches improve model performance and alleviate the computational burden and biases associated with traditional feature extraction methods in ML models.

Both approaches significantly contribute to helping the clinical community in predicting the mental health of social media users without attaching any social stigma. However, neither of these models achieves the accuracy of a human moderator. These methods have limitations, including the fact that the majority of deep learning networks fail to capture domain knowledge because they are trained using general-purpose text data. On the other hand, machine learning-based methods struggle to generalize on unseen data.

	Not depression	Moderate	Severe
Train	2,755	3,678	768
Test	848	2,169	228
Total	**3,603**	**5,847**	**996**

Table 1: Dataset distribution

The proposed model utilizes transformer-based Large Language models like PubMed BERT to generate feature vectors to address these issues. This approach allows the model to capture domain knowledge and context information from social media text, enabling it to predict depression levels more effectively. By incorporating these improvements, the proposed model aims to enhance accuracy and better understand users' mental health.

2.3 Task and dataset description

The current study utilizes the dataset from the LT-EDI@RANLP 2023 shared task (S et al., 2022), which focuses on detecting signs of depression in a social media text. The organizers of the shared task have provided a challenge regarding the identification of depression levels in English social media comments. The dataset consists of a text field and a label field, with the labels being ”not depression,” ”moderate,” and ”severe.” According to the organizers, the data was collected from YouTube comments (S et al., 2022). The detailed distribution of the dataset is presented in Table 1. However, the dataset is highly skewed, with the majority of the comments labelled as ”moderate” and very few instances of ”severe” comments.

3 Model building

In this section, we outline the model submitted for the shared task of identifying depression levels in social media data. The proposed model comprises three primary steps: data cleaning and preprocessing, feature extraction, and classification. This section will thoroughly explain each of these stages. The architecture of the model is illustrated in Fig 1.

3.1 Data pre-processing

According to the shared task organizers, data has been collected from YouTube comments. As social media data often contains noise, certain steps have been taken to clean the data. The text data includes punctuation, hyperlinks, URLs, stop words, and

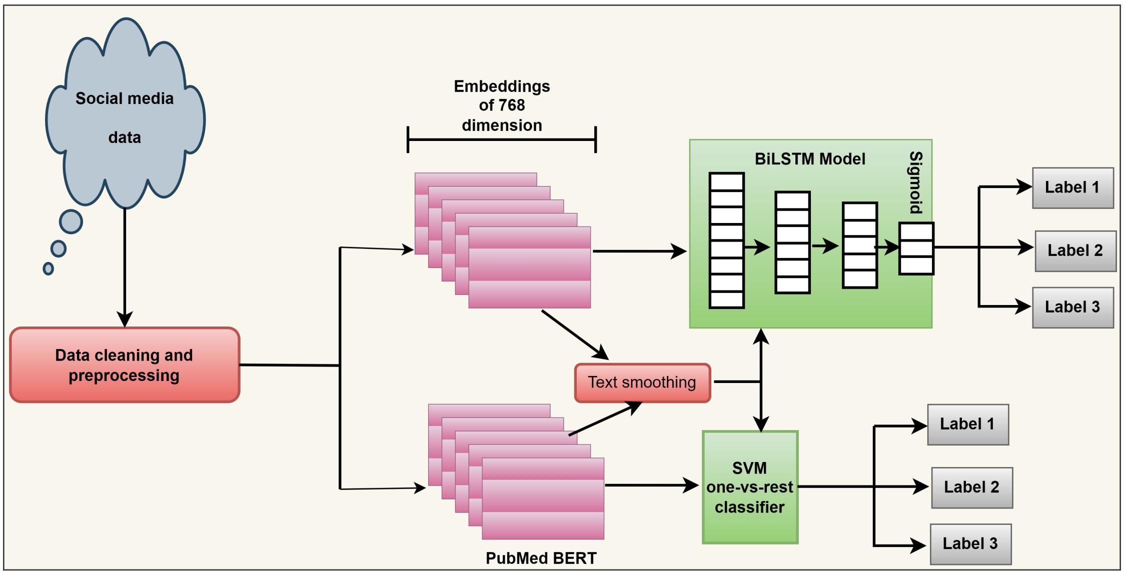

Figure 1: Experimented model architecture

Model	Hyper-parameter

SVM	Kernal = 'linear'
	C = 1

BiLSTM	Node size = 60,30,10
	Drop out rate = 0.5
	Loss = 'categorical_crossentopy'
	Optimiser = 'Adam'
	Batch_size = 100
	Epochs = 10

Table 2: Hyper-parameter

numerical data, which do not contribute to the final class prediction. To address this, we have removed these elements using simple string operations. Stop words have been eliminated using the NLTK library. Additionally, the text has been converted to lowercase to avoid token redundancy. Finally, lemmatization has been applied to convert social media slang to its root words. All of these steps have been performed using the NLTK toolkit [3].

After the pre-treatment, the data is subjected to tokenization, where we apply the BERT tokenizer to convert the text input into tokens. Subsequently, padding is performed on the tokenized data to ensure all comments possess a fixed-length sequence. Finally, masking is applied to the padded sequence to eliminate the influence of padded tokens on label prediction.

3.2 Feature Extraction

The proposed model utilizes a pre-trained language model called PubMed BERT, obtained from the Hugging Face library [4]. PubMed BERT is a variant of the original BERT model, trained on clinical data (Gu et al., 2020). Its architecture closely resembles that of the original BERT model(Kenton and Toutanova, 2019). The main objective of the feature extraction process in this model is to represent high-dimensional text data into lower-dimensional embedding vectors. To achieve this, padded and masked sequences are provided as input. The model extracts the embeddings from the [CLS] token to generate the embedding vectors. This token represents the entire sentence and provides a bidirectional representation of the input text. By utilizing the embeddings from the [CLS] token, the model captures the overall semantic meaning of the text. The advantage of employing PubMed BERT in the proposed model lies in its training on clinical data, which enables it to incorporate domain-specific information into the embeddings. This makes the model well-suited for tasks involving depression analysis. After obtaining the embeddings from PubMed BERT, they are passed as input to the data augmentation and classification stage.

To address the issue of highly skewed data towards the moderate label, the proposed method incorporates text smoothing on input embeddings to achieve a more balanced representation of the

[3] https://www.nltk.org/

[4] https://huggingface.co/

Model	F1-moderate	F1-not depression	F1-severe	Macro-F1
PubMed with BiLSTM (without smoothing)	0.66	0.30	0.10	0.41
PubMed with BiLSTM (with smoothing)	0.20	0.37	0.64	0.45
PubMed with SVM (without smoothing)	0.70	0.40	0.33	0.48
PubMed with SVM (with smoothing)	**0.44**	**0.55**	**0.66**	**0.54**

Table 3: Comparative results of the proposed model

overall input text sequences. Subsequently, the balanced data is fed as input to the classification layer. The proposed method conducts experiments using both the balanced and original text data, and the results and discussion section presents the findings of these experiments.

3.3 Classification

The primary objective of the classification stage is to convert the input embeddings into corresponding depression levels. To achieve this, the proposed model experimented with different machine learning and deep learning-based models.

The proposed method utilized the Support Vector Machine (SVM) classifier among the machine learning-based models. Since the problem involves multiclass classification, the proposed method employed the One-Vs-Rest classifier from SVM to identify depression levels in a social media text. Further, the proposed method also experimented with a Bidirectional Long Short-Term Memory (BiLSTM) model. The BiLSTM model was constructed using two BiLSTM layers with 60 and 30 neurons, and a dense layer with ten units was added after BiLSTM layers, and a dropout rate of 0.5 was applied, indicating that 50% of the input units were randomly dropped out. Finally, the output layer consisted of a dense layer with three units representing the number of classes, and the softmax activation function was added to predict the output class. The hyperparameters used to train both models are illustrated in Table 2. These hyperparameter values were selected based on experimental trials. The input for the classification stage was taken from PubMed BERT embeddings, which have a vector dimension of 768. Implementation details of the proposed model can be found in the GitHub repository [5].

Team name	Macro-F1	Rank
DeepLearningBasil	0.47	1
DeepBlueAI	0.446	2
Cordyeeps_ssl	0.441	3
iicteam	0.439	4
CIMAT-NLP	0.439	5
IIITDWD	**0.359**	**21**

Table 4: Top performing teams

4 Result and Discussion

The proposed model was trained to identify class labels such as ’not depression,’ ’moderate,’ and ’severe.’ The comparative results of the model are summarized in Table 3.

The proposed model was tested on both balanced data after smoothing and the original text to assess the impact of text smoothing on its performance. As shown in Table 3, the model exhibited significant performance in predicting the ’moderate’ label before smoothing. However, its performance with the other two class labels was moderate, resulting in a reduced macro-F1 score. Text smoothing resulted in a more evenly distributed weighted F1 score across all labels.

In evaluating the model performance for the shared task LT-EDI @RANLP2023, the organizers utilized the macro-F1 score. Among the proposed methods, the combination of PubMed BERT with SVM on balanced data achieved a higher macro-F1 score and was therefore chosen for submission in the shared task. Our proposed model received the 21st rank among the participating teams, with a macro-F1 score of 0.36 on unseen data. Table 4 displays some of the top-performing teams in the competition, including our proposed model (highlighted in bold).

[5] https://github.com/shankarb14/RANLP-2023

5 Conclusion and future enhancements

The study presents the model submitted to the LT-EDI@RANLP 2023 shared task, which aims to detect signs of depression in a social media text. The proposed model experimented with two approaches: SVM and BiLSTM as classifiers. The study concludes that PubMED BERT embeddings combined with SVM classifier on a balanced dataset achieve more uniformly distributed weighted F1 scores across all the labels. The proposed model secured the 21st rank in the competition. However, the model's performance could be further improved by developing a more robust algorithm capable of capturing domain-specific information and contextual details from the input text.

References

Mohammad Akbari, Xia Hu, Nie Liqiang, and Tat-Seng Chua. 2016. From tweets to wellness: Wellness event detection from twitter streams. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Shankar Biradar and Sunil Saumya. 2022. Iiitdwd@ tamilnlp-acl2022: Transformer-based approach to classify abusive content in dravidian code-mixed text. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 100–104.

Shankar Biradar, Sunil Saumya, and Arun Chauhan. 2021. Hate or non-hate: Translation based hate speech identification in code-mixed hinglish data set. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 2470–2475. IEEE.

Shankar Biradar, Sunil Saumya, and Arun Chauhan. 2022. Fighting hate speech from bilingual hinglish speaker's perspective, a transformer-and translation-based approach. *Social Network Analysis and Mining*, 12(1):87.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, Subalalitha Cn, John Philip McCrae, Miguel Ángel García, Salud María Jiménez-Zafra, Rafael Valencia-García, Prasanna Kumaresan, Rahul Ponnusamy, et al. 2022. Overview of the shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 378–388.

Xuetong Chen, Martin D Sykora, Thomas W Jackson, and Suzanne Elayan. 2018. What about mood swings: Identifying depression on twitter with temporal measures of emotions. In *Companion proceedings of the the web conference 2018*, pages 1653–1660.

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.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

Sharath Chandra Guntuku, Anneke Buffone, Kokil Jaidka, Johannes C Eichstaedt, and Lyle H Ungar. 2019. Understanding and measuring psychological stress using social media. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 214–225.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818.

S Kayalvizhi, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, et al. 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*, pages 331–338.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Jina Kim, Jieon Lee, Eunil Park, and Jinyoung Han. 2020. A deep learning model for detecting mental illness from user content on social media. *Scientific reports*, 10(1):1–6.

Harnain Kour and Manoj K Gupta. 2022. An hybrid deep learning approach for depression prediction from user tweets using feature-rich cnn and bidirectional lstm. *Multimedia Tools and Applications*, 81(17):23649–23685.

Huijie Lin, Jia Jia, Liqiang Nie, Guangyao Shen, and Tat-Seng Chua. 2016. What does social media say about your stress?. In *IJCAI*, pages 3775–3781.

Jingfang Liu and Mengshi Shi. 2022. A hybrid feature selection and ensemble approach to identify depressed users in online social media. *Frontiers in Psychology*, 12:6571.

Inna Pirina and Çağrı Çöltekin. 2018. Identifying depression on reddit: The effect of training data. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 9–12.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. E-bert: Efficient-yet-effective entity embeddings for bert. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 803–818.

Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. Iiitt@ lt-edi-eacl2021-hope speech detection: there is always hope in transformers. *arXiv preprint arXiv:2104.09066*.

Kayalvizhi S, 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*, Dublin, Ireland. Association for Computational Linguistics.

Sunil Saumya Shankar Biradar and Arun Chauhan. 2021. mbert based model for identification of offensive content in south indian languages. In *Working Notes of FIRE 2021-Forum for Information Retrieval Evaluation (Online). CEUR*.

Michael M Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7:44883–44893.

Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. 2015. Recognizing depression from twitter activity. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 3187–3196.

Mudasir Ahmad Wani, Mohammad A ELAffendi, Kashish Ara Shakil, Ali Shariq Imran, and Ahmed A Abd El-Latif. 2022. Depression screening in humans with ai and deep learning techniques. *IEEE Transactions on Computational Social Systems*.

Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. 2023. Hierarchical convolutional attention network for depression detection on social media and its impact during pandemic. *IEEE Journal of Biomedical and Health Informatics*.