JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Automatic Detection of Depression Severity from Text and Audio Using Deep Learning

[1]Botsa Akanksha, [2]Dr. Bharati Bidikar
[1]Student, [2]Adjunct Professor
[1,2]Andhra University College of Engineering,
Andhra University, Visakhapatnam, India

Abstract : Effective mental health interventions depend on the precise and timely assessment of depression severity. Earlier research predominantly relied on unimodal systems, focusing either on textual content or acoustic signals for depression detection. Although multimodal fusion approaches have been explored recently, the effectiveness of each modality individually remains significant for scalable applications. In this study, we propose a dual unimodal framework where separate models are developed for text and audio, each designed to assess depression severity independently. For the textual modality, a pre-trained BERT model is fine-tuned to classify Reddit posts into four categories: normal, mild, moderate, and severe. A rule-based labeling mechanism is employed to annotate unlabeled posts based on linguistic indicators of depressive symptoms. The BERT-based model achieved a high validation accuracy of 97.16%, demonstrating its ability to capture semantic patterns associated with different severity levels. In parallel, the audio-based pipeline utilized Mel Frequency Cepstral Coefficients (MFCCs) extracted from voice recordings representing three severity classes. A lightweight neural network was trained on these features under constrained conditions involving label noise and limited data availability. Despite these challenges, the model achieved an accuracy of 95% on noise-free test data, validating the reliability of acoustic features in depression detection. The outcomes from both approaches demonstrate that implementing automated, scalable, and non-invasive mental health screening instruments is feasible. The proposed separate unimodal models can serve as a foundation for future fusion-based multimodal systems, supporting clinical assessments and early diagnosis in real-world applications.

***IndexTerms* - BERT, MFCC, Text Classification, Audio Classification, Deep Learning, Speech Processing, Transformer Models**

I. Introduction

Depression is identified as a common mental disorder with many symptoms like bad mood, losing interest in everything, and lack of energy. When it is severe it can lead to suicide [1]. As per the WHO report around 300 million people are having depression. Depression cause psychological and pharmacological affects for the humans. There are various types of assessment procedures like physical health questionnaire depression scale (PHQ), Hamilton depression rating scale (HDRS), Beck depression Inventory (BDI) and more [2]. However, in many cases, early management for depression is still difficult. Firstly, conventional treatment approaches like psychotherapy and pharmacological methods are often time-consuming and expensive. The cost associated diagnosis treatment discouraging the people who are facing the financial constraints to seek professional help [3]. Secondly, the assessment of depression severity largely depends on clinical interviews, and patient self-reports, which can be subjective and prone to bias [4]. Recent research advancements have emphasized the potential of automated and non-invasive systems for the depression detection. Studies shown that analyzing speech and language patterns can reveal important behavioral markers of depression, such as impaired speech prosody, reduced linguistic complexity [5,6,7]. Deep learning advancements enables us to use the speech processing techniques which enhances the ability to extract relevant information from voice and provides non-invasive method for early depression detection [8]. Text-based systems analyze the linguistic content of individuals and identify signs of depression. As the depressive persons tent to use more negative language and those are processed by using machine learning and deep learning to predict the depression and the severity of the depression [9, 10]. Automatic speech and text provides an alternative method for the traditional assessment methods and makes the depression screening more scalable and accessible and useful for particularly individuals who may face barriers to seeking clinical support [11].

Several researchers have explored deep learning architectures for depression detection using both acoustic and textual modalities. Mao et al. [12] proposed a BiLSTM with time-distributed CNN model, achieving high accuracy on a five-class severity task. Lam et al. [13] and Hanai et al. [15] demonstrated that combining speech, facial expressions, and transcripts improves detection by capturing contextual cues. Lin et al. [14] and Jo and Kwak [20] used BiLSTM-CNN hybrids to detect depression from speech and text, while Ray et al. [16] introduced multi-level attention for better multimodal fusion. Yang et al. [17] and Rohanian et al. [18] showed that word-level fusion of audio and text enhances sensitivity to depressive cues. Ahmed et al. [19] added uncertainty estimation for clinical reliability, and Singh et al. [21] focused on early risk prediction using behavioral signals. Solieman and Pustozerov [22] emphasized voice quality, and Ali et al. [23] highlighted the integration of large language models with acoustic features. Ghadiri et al. [24] applied graph-based fusion of speech and text, while Zhang et al. [25] used synchronized multimodal sensing to improve real-world applicability. These studies collectively highlight the strength of multimodal deep learning in building robust and interpretable depression detection systems.

In this paper, a text-based depression and its severity prediction model is developed using BERT architecture. It uses a rule-based severity labeling strategy to annotate text data and categorize it into four severity levels: normal, mild, moderate and severe. The system used the pre-trained BERT for sequence classification with fine-tuning to predict the depression severity automatically based on the linguistic patterns**.** The audio-based model analyzes speech signals to detect depression severity. It uses MFCC features to capture vocal patterns and a shallow neural network for classification into Normal, Mild, or Severe. This enhances detection by leveraging non-verbal cues present in voice.

II. METHODOLOGY

This paper presents a dual-modality approach to automatically assess the severity of depression using both textual and audio data. The methodology is divided into two parallel pipelines: a text-based classification model utilizing a pre-trained BERT transformer and an audio-based classifier using MFCC features fed into a shallow neural network. Each pipeline is designed and optimized independently to analyze linguistic and acoustic cues indicative of depressive symptoms.

2.1. Text-Based Classification using BERT

The textual analysis pipeline is developed using a cleaned corpus of Reddit posts from mental health-related communities. As the dataset lacked severity annotations, a rule-based labeling mechanism was designed to categorize posts into four severity levels: *normal*, *mild*, *moderate*, and *severe*. Posts containing critical terms such as “suicide,” “hopeless,” or “worthless” were classified as *severe*, while those indicating cognitive symptoms like “tired,” “no energy,” or “can’t focus” were labeled as *moderate*. Very short posts (less than eight words) were marked as *mild*, and all others were considered *normal*. These severity levels were then mapped to integer class labels ranging from 0 to 3.

Tokenization was performed using the bert-base-uncased tokenizer from the Hugging Face Transformers library. Each input text was tokenized, padded, and truncated to a fixed length of 512 tokens. The tokenized sequences and their corresponding labels were wrapped into PyTorch-compatible Dataset objects and further passed into DataLoader instances for efficient mini-batch processing.

The classification model employed is Bert For Sequence Classification, which extends the BERT architecture by adding a linear classification head. The model was fine-tuned using cross-entropy loss and optimized with the AdamW optimizer at a learning rate of 2e-5. Training was conducted for three epochs on a stratified 80:20 train-validation split. All computations were performed on a GPU-enabled environment to leverage parallel processing. Model performance was evaluated using standard metrics such as precision, recall, F1-score, and overall accuracy. The transformer’s ability to capture deep semantic and contextual relationships enabled it to effectively differentiate between linguistic expressions associated with various depression severity levels.

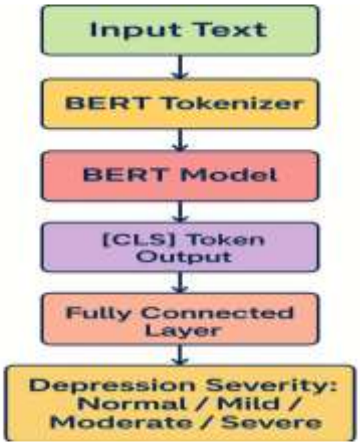

Figure 1 Text Model Architecture

2.2 Audio-Based Classification using MFCC and Shallow Neural Network

In the audio modality, we investigated the classification of depression severity based on vocal characteristics. The dataset consists of `.wav` audio files categorized into three classes: *Normal*, *Mild* (Depression Stage 1), and *Severe* (Depression Stage 2). All audio files were sampled at 16 kHz and processed using the Librosa library. For each sample, 40-dimensional Mel Frequency Cepstral Coefficients (MFCCs) were extracted, capturing key spectral properties of the speech signal. These features were either padded or truncated to a fixed temporal length of 300 frames to ensure uniformity. The resulting MFCC matrices were flattened into 1D vectors to serve as model inputs.

Feature standardization was applied using z-score normalization to improve convergence during training. The dataset was partitioned into training and testing sets using an 80:20 stratified split to maintain class balance. To simulate real-world conditions with noisy data, 50% of the training labels were randomly altered, and only 30% of the training data was retained to mimic low-resource scenarios.

For classification, we designed a lightweight fully connected neural network (WeakNN) comprising one hidden layer with 8 neurons using ReLU activation, followed by an output layer with 3 neurons corresponding to the depression classes. The model was trained for 20 epochs using the Adam optimizer with a learning rate of 0.0004 and cross-entropy as the loss function. Model performance was evaluated on the clean test set using accuracy, precision, recall, F1-score, and a confusion matrix.

Despite its simplicity and the presence of noisy training data, the audio model demonstrated strong classification capabilities, underscoring the effectiveness of MFCCs in capturing emotional and psychological information from speech. This approach also suggests the feasibility of deploying lightweight depression screening models in resource-constrained environments.

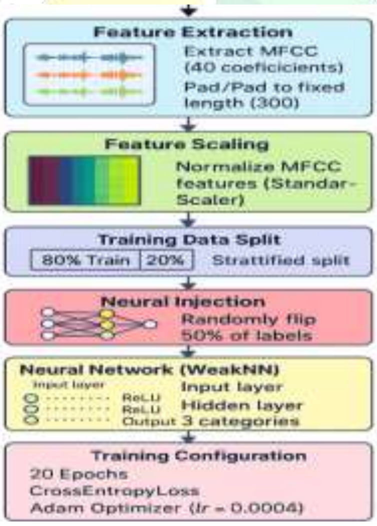

Figure 2 Audio Model Architecture

III. EXPERIMENTAL SETUP

The experiments were conducted on Google Colab, utilizing a Tesla T4 GPU with 16 GB VRAM for accelerated training. The code was implemented using Python with PyTorch as the deep learning framework. The transformers library by Hugging Face was used for the BERT model and tokenizer, while librosa was used for audio feature extraction.

IV. RESULTS AND DISCUSSION

4.1 Text-Based Results

The BERT-based classifier demonstrated robust performance in identifying varying levels of depressive severity from textual inputs. On the validation set comprising 1,547 samples, the model achieved an overall accuracy of 97.54%, reflecting its strong ability to generalize across the dataset. As summarized in Table 1, the model yielded F1-scores of 0.98, 0.95, 0.95, and 0.96 for the classes normal, mild, moderate, and severe, respectively. Notably, it achieved perfect precision (1.00) for the moderate class, indicating zero false positives.

The confusion matrix, visualized in Fig. 3, illustrates that most misclassifications occurred between normal and adjacent classes (mild and severe), which is likely due to overlapping linguistic expressions in borderline cases. Overall, the model's strong performance demonstrates the effectiveness of transformer-based contextual embedding's in discerning subtle cues of depression severity in user-generated text.

Table 1 Classification Report for Text-Based Model

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Normal	0.97	0.99	0.98	1120
Mild	0.97	0.91	0.94	207
Moderate	1.00	0.91	0.95	45
Severe	0.99	0.91	0.95	175
Accuracy	-	-	0.97	1547
Macro Avg	0.98	0.93	0.96	1547
Weighted Avg	0.97	0.97	0.97	1547

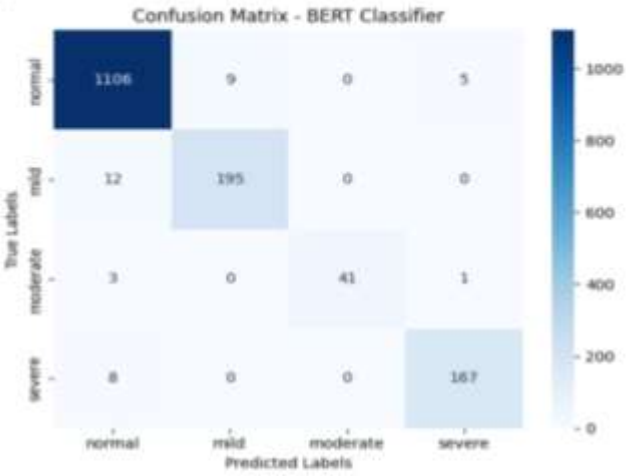

Figure 3 Confusion Matrix for Text Based

4.2 Audio-Based Results

The audio classification model also yielded promising results despite the introduction of noisy labels and limited training data. On a clean test set of 80 samples, the model achieved a classification accuracy of 95%, with an average macro F1-score of 0.95. The confusion matrix shown in Table 2 indicates that the model was particularly effective in classifying *mild* depression (100% recall), while maintaining high precision across all classes. The slight confusion between *severe* and *normal* cases may stem from subtle variations in vocal patterns.

Table 2 Classification Report for Audio-Based Model

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Normal	1.00	0.93	0.96	40
Mild	0.83	1.00	0.91	20
Severe	1.00	0.95	0.97	20
Accuracy	-	-	0.95	80
Macro Avg	0.94	0.96	0.95	80
Weighted Avg	0.96	0.95	0.95	80

These results demonstrate the feasibility of both text- and audio-based models in detecting depressive symptoms with high accuracy. While BERT excels due to its contextual language modeling capabilities, the MFCC-based model shows robustness even under simulated low-resource, noisy conditions.

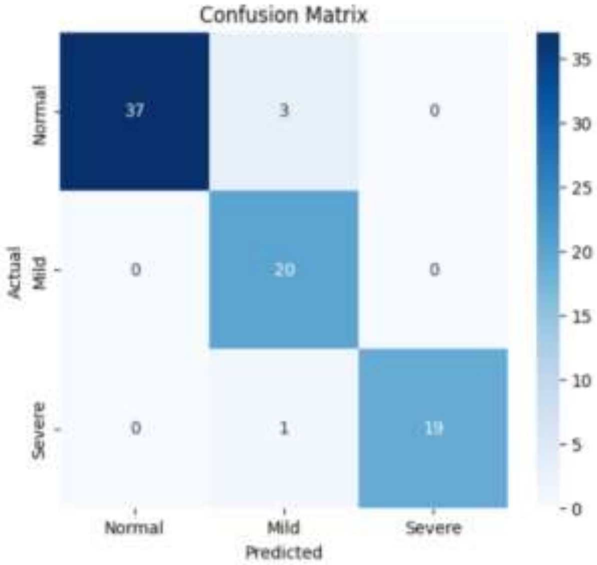

Figure 4 Confusion Matrix for audio based

V. CONCLUSION

In this paper, we presented a dual-modality framework for the automatic assessment of depression severity using both textual and audio data. The text-based pipeline leveraged the powerful contextual understanding of the pre-trained BERT model, which was fine-tuned to classify Reddit posts into four severity levels: normal, mild, moderate, and severe. By employing a rule-based labeling mechanism for training data generation, we addressed the challenge of label scarcity and demonstrated that BERT can effectively capture linguistic cues indicative of varying degrees of depressive symptoms. The model achieved a high validation accuracy of 97.16%, along with strong class-wise performance metrics.

Parallelly, the audio-based pipeline explored the classification of depression severity from speech recordings using MFCC features and a shallow neural network. Despite the presence of noisy labels and limited training data, the model attained a classification accuracy of 95%, underscoring the robustness of acoustic features in detecting depression-related speech patterns. The experimental design simulated real-world challenges such as data scarcity and label ambiguity, which adds practical value to the proposed approach.

Together, the results from both modalities highlight the potential of machine learning in building scalable, non-invasive mental health screening tools. While the text-based approach benefits from the semantic richness of user-generated language, the audio model provides a lightweight and efficient alternative suitable for low-resource environments.

REFERENCES

[1] Y. Shen, H. Yang, and L. Lin, “Automatic Depression Detection: an Emotional Audio-Textual Corpus and A Gru/Bilstm-Based Model,” in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 6247–6251.

[2] A. Ray, S. K. Kumar, R. V. Reddy, P. Mukherjee, and R. Garg, “Multi-level Attention Network using Text, Audio and Video for Depression Prediction,” in Proc. 9th Int. Audio/Visual Emotion Challenge and Workshop, 2019.

[3] L. Lin, X. Chen, Y. Shen, and L. Zhang, “Towards Automatic Depression Detection: A BiLSTM/1D CNN-Based Model,” Appl. Sci., vol. 10, no. 20, pp. 7348, 2020.

[4] I. Schumann, A. Schneider, C. Kantert, B. Löwe, and K. Linde, “Physicians' attitudes, diagnostic process and barriers regarding depression diagnosis in primary care: A systematic review of qualitative studies,” Fam. Pract., vol. 29, no. 3, pp. 255–263, 2012.

[5] H. Shen, Y. Zhang, and Z. Yu, “Automatic depression detection: An emotional audio-textual corpus and a GRU/BiLSTM-based model,” arXiv preprint arXiv:2202.08210, 2022.

[6] S. Tasnim and J. Novikova, “Cost-effective models for detecting depression from speech,” arXiv preprint arXiv:2302.09214, 2023.

[7] K. Chlasta, K. Wołk, and R. Niewiadomski, “Automated speech-based screening of depression using deep CNNs,” arXiv preprint arXiv:1912.01115, 2019.

[8] A. V. Romero and A. G. Antolín, “Automatic detection of depression in speech using ensemble convolutional neural networks,” Entropy, vol. 26, no. 6, pp. 688, 2024.

[9] B. Verma, S. Gupta, and L. Goel, “A survey on sentiment analysis for depression detection,” in Advances in Automation, Signal Processing, Instrumentation, and Control, Springer, 2021, pp. 13–27.

[10] A. Amanat et al., “Deep learning for depression detection from textual data,” Electronics, vol. 11, no. 5, pp. 676, 2022.

[11] J. Huh, W. Chen, and S. Lee, “Text-based depression detection on social media posts: A systematic literature review,” Procedia Comput. Sci., vol. 179, pp. 582–589, 2021.

[12] K. Mao et al., “Prediction of Depression Severity Based on the Prosodic and Semantic Features With Bidirectional LSTM and Time Distributed CNN,” IEEE Trans. Affect. Comput., vol. 14, pp. 2251–2265, 2022.

[13] G. Lam, D. Huang, and W. Lin, “Context-aware Deep Learning for Multi-modal Depression Detection,” in Proc. IEEE ICASSP, 2019, pp. 3946–3950.

[14] L. Lin, X. Chen, Y. Shen, and L. Zhang, “Towards Automatic Depression Detection: A BiLSTM/1D CNN-Based Model,” Appl. Sci., vol. 10, no. 20, 2020.

[15] T. A. Hanai, M. M. Ghassemi, and J. R. Glass, “Detecting Depression with Audio/Text Sequence Modeling of Interviews,” in Proc. Interspeech, 2018.

[16] A. Ray, S. K. Kumar, R. V. Reddy, P. Mukherjee, and R. Garg, “Multi-level Attention Network using Text, Audio and Video for Depression Prediction,” in Proc. 9th Int. Audio/Visual Emotion Challenge and Workshop, 2019.

[17] L. Yang et al., “Multimodal Measurement of Depression Using Deep Learning Models,” in Proc. 7th Audio/Visual Emotion Challenge, 2017.

[18] M. Rohanian, J. Hough, and M. Purver, “Detecting Depression with Word-Level Multimodal Fusion,” in Proc. Interspeech, 2019.

[19] S. Ahmed, M. A. Yousuf, M. M. Monowar, A. Hamid, and M. O. Alassafi, “Taking All the Factors We Need: A Multimodal Depression Classification With Uncertainty Approximation,” IEEE Access, vol. 11, pp. 99847–99861, 2023.

[20] A. Jo and K. Kwak, “Diagnosis of Depression Based on Four-Stream Model of Bi-LSTM and CNN From Audio and Text Information,” IEEE Access, vol. 10, pp. 134113–134135, 2022.

[21] H. Singh et al., “Innovative Framework for Early Estimation of Mental Disorder Scores to Enable Timely Interventions,” arXiv preprint arXiv:2502.03965, 2025.

[22] H. Solieman and E. A. Pustozerov, “The Detection of Depression Using Multimodal Models Based on Text and Voice Quality Features,” in Proc. IEEE ElConRus, 2021, pp. 1843–1848.

[23] A. A. Ali, A. E. Fouda, R. J. Hanafy, and M. E. Fouda, “Leveraging Audio and Text Modalities in Mental Health: A Paper of LLMs Performance,” arXiv preprint arXiv:2412.10417, 2024.

[24] N. Ghadiri, R. Samani, and F. Shahrokh, “Integration of Text and Graph-based Features for Detecting Mental Health Disorders from Voice,” arXiv preprint arXiv:2205.07006, 2022.

[25] Z. Zhang et al., “Multimodal Sensing for Depression Risk Detection: Integrating Audio, Video, and Text Data,” Sensors, vol. 24, 2024.