Multimodal Mental Health Detection using BERT and Audio Analysis

Aparna N[1], Suresha D[2], Spoorthi B[3]
[1]*M. Tech Student, Dept. of Computer Science and Engineering, Srinivas Institute of Technology, Mangaluru , 671124, India*
[2] *Professor, Dept. of Computer Science and Engineering, Srinivas Institute of Technology, Mangaluru, 671124, India*
[3]*Assistant Professor, Dept. of Computer Science and Engineering, Srinivas Institute of Technology, Mangaluru, 671124, India*
doi.org/10.64643/IJIRTV12I12-206753-459

***Abstract*—There have been cases of the increasing number of mental illnesses like depression and suicidal thoughts, and there is an urgent need for early diagnosis and treatment. In this research, the researchers have developed a unique multimodal system to determine one's mental status based on text analysis and voice recognition. The text recognition process has been done using fine-tuning of a BERT model, while Mel Frequency Cepstral Coefficients (MFCC) have been used to process the voice input. As for the results, the proposed system identifies the user input into one of the three categories, including Normal, Depressed, and High Suicide Risk, along with a corresponding confidence score and the risk value assigned to each input. The text-based model yields 94% accuracy, while the voice model attains 95% accuracy. As it can be seen, the integration of both approaches helps get better results, as linguistic context and emotions can be used for further analysis. It is also worth mentioning that the system has been implemented in real-time using the Streamlit web application.**

***Index Terms*—Mental Health Detection, BERT, Multimodal Learning, Emotion Recognition, Natural Language Processing, Suicide Risk Prediction.**

I. INTRODUCTION

Mental well-being is becoming more and more relevant in contemporary society, with many people struggling to cope with academic pressures, work-related expectations, and personal problems. Thus, one may worry not only about their physical health but also about their mental well-being. Words like stress, anxiety, and depression have become part of our daily vocabulary. It may sometimes become quite difficult to draw a line between natural emotions that people tend to experience in life and mental conditions, with people confusing themselves. In addition, existing methodologies used in recognizing such conditions usually employ techniques like keyword matching and machine learning algorithms.

The problem with these methodologies is that they cannot analyze the context adequately and may give erroneous results. For example, the experience of temporary frustration may be interpreted as a mental condition when in fact, it is not one.

The latest achievements in the domain of artificial intelligence open up possibilities to design even more sophisticated tools to assess the user's psychological condition based on the analysis of different parameters. Specifically, we propose to develop a multi-modal system aimed at evaluating mental condition through the interaction of the user with other people. Text processing will be conducted using a fine-tuned BERT model, allowing for obtaining deep insight into the nature of the information presented by users in their messages.

Moreover, our solution includes audio analysis that allows for extracting features from audio files and identifying specific emotions. All the results obtained from the analysis will be conveniently displayed in one place. In any case, we would like to highlight that our tool cannot be used as a replacement for professional diagnostic methods.

Our main purpose is to raise user's awareness about his or her mental condition.

II. RELATED WORK

The rising occurrence of psychological disorders has necessitated the development of automated diagnosis techniques based on machine learning and deep learning algorithms. At the beginning, the existing techniques mainly relied on NLP methodologies along with the use of classification algorithms including Support Vector Machine (SVM) and Naive Bayes algorithm. Although these techniques yielded fairly good results for basic sentiment analyses, they proved to be insufficient in dealing with complicated linguistic nuances such as sarcasm, negations, and contradictions [1]. As a result of introducing deep learning in this field, new models were devised for analysing textual information. In this regard, models like RNN and LSTM showed increased effectiveness in identifying depressive trends along with emotions contained in text documents[2]. The existing models are not very good at understanding relationships between things that're far apart in a text. They also do not do a job of comprehending the entire context of long texts. The existing models still have a problem with this. They just cannot capture the relationships between things that're distant from each other in a text. The existing models need to be better, at understanding the context of texts. The BERT architecture, which uses the transformer model, has changed the game of language processing by being capable of processing both the backward and forward context of words, including both the past and future words of a certain text. Hence, BERT proved itself to be much more effective in identifying language nuances, becoming perfect for detecting the symptoms of mental illnesses. In multiple papers, it has shown its efficiency in recognizing suicidal intentions and depression among other problems [3][4].

Other detection methods include those based on multimodal systems where apart from analysing text, audio is also analysed. Such systems enable the tone of the voice to be detected in addition to text analysis. The audio can be analysed by applying different algorithms such as MFCC algorithm. These algorithms can detect emotions in the text [5]. Text and speech go hand-in-hand in this case since both can give information about the contents and how they are delivered [9][10].

Despite these enhancements, there still exist certain weaknesses in these models, particularly with regard to minimizing the occurrence of false positives and the understanding of indirect verbal cues such as denials. This means that these models need to be further developed to enable them to effectively combine the auditory and linguistic elements [6][7][8].

III. SYSTEM OVERVIEW

The platform represents a multimodal mental well-being assessment tool. It analyzes both textual information and voice messages to determine the mood of a person. Textual analysis assists in detecting emotional conditions through text. Audio analysis includes tone, pitch, and other features of speech. Integration of both methods increases the quality of results received. Natural language processing and audio signal processing technologies are used in evaluating both texts and voices correspondingly. It allows obtaining more precise insights into mental well-being. The web application is developed with the help of Streamlit. Its user-friendly interface allows users to input textual data or upload voice files.

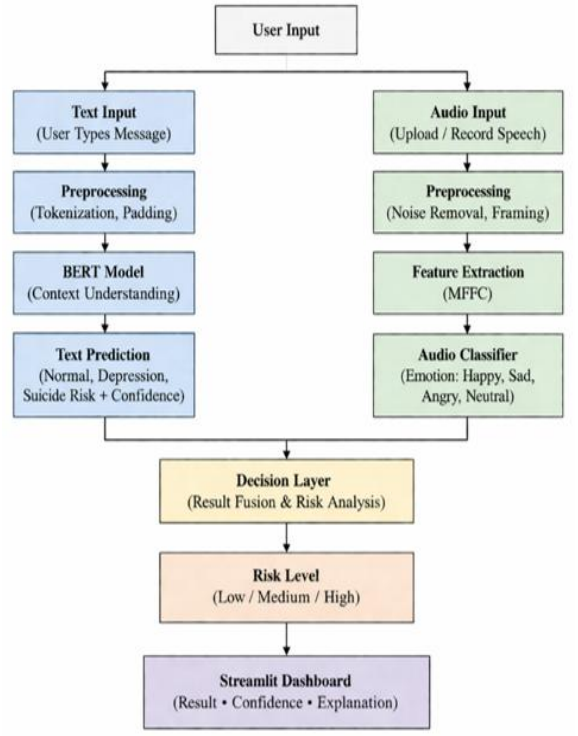

Figure 1: System Architecture

A. Input Module

The user can choose to input his ideas either through typing or by using voice commands, thereby making things easier for the user in expressing himself. The merging of the two will aid the user in conveying his emotions better.

B. Text Processing and Classification

The text undergoes a series of processes prior to the actual analysis. This includes deleting any unnecessary details, standardizing the text format to maintain consistency, and dividing the text into smaller parts. After that, it passes through a highly advanced BERT architecture that has undergone specific training for emotion or mental well-being detection purposes. Unlike other approaches, the analysis performed on the text also considers its context within the sentence. Subsequently, the system classifies the text under one of the following categories: Normal, Depression, or Suicide Risk.

C. Audio Processing and Emotion Detection

The features of audio are obtained through Mel Frequency Cepstral Coefficients (MFCC). The features of the audio are then classified through a machine learning algorithm where emotions such as happiness, sadness, or anger are identified in the voice signal. Finally, the identified emotions are interpreted to give an indication of the mental state.

D. Decision Fusion Layer

In this module, both the results from the text classification process and the audio analysis process contribute to the final decision on whether a person is mentally stable or not. This integration of both sets of information ensures that the decision made will be accurate and without any doubts.

E. Risk Assessment and Output

Once all the results from the analysis have been combined, the system is then able to determine the level of risk that is involved with the project and classify the same as being either Low, Medium, or High risk. This result is shown together with the forecast , including the probability and a brief description for clarity purposes.

IV. IMPLEMENTATION

The system will be designed using an approach that involves the integration of deep learning methods along with traditional machine learning methods, and advanced web technologies. The steps for implementing the system will involve gathering and organizing appropriate data sets, creating and training prediction models, and incorporating the models within the interface of a user-friendly dashboard.

A. Dataset Preparation

The proposed model is trained with the help of three different sets of data, out of which two belong to the textual sources and one belongs to the audio source. The textual dataset is made up of data from datasets related to suicidal and depressive behaviour. The suicidal behaviour dataset consists of data from users who are suicidal or not.

On the other hand, the depressive behaviour dataset comprises data that is marked based on the symptoms of depression. Once the data from both datasets is collected, it is segregated into three groups: normal, depression, and suicide.

For improving the data quality, there are a few pre-processing techniques that are considered to be useful. Some of the techniques are the removal of null values, duplicate values, and irrelevant columns. Moreover, balancing is another technique where the same number of entries are selected from all categories. This technique helps in avoiding any form of bias by the model towards a specific category. As a result, the accuracy improves significantly.

The audio data utilized in the model contains recorded speech emotions, which are marked with four kinds of emotions like happiness, sadness, anger, and neutrality. These emotions are related to mental illnesses.

Table 1: Statistical Summary of the Suicide Dataset

Metric	Value
Samples	232,074
Avg Length	132 words
Balance	50 / 50

Table 2: Statistical Summary of the Depression Dataset

Metric	Value
Samples	7,873
Avg Length	73 words
Balance	50.3 / 49.7

Figure 3: Audio dataset class distribution

B. Text Classification Module

Pretrained BERT model is utilized for classification tasks. Firstly, the input text goes through the transformation process by applying the BERT tokenizer. Here, the input text is encoded into numbers without exceeding a particular maximum number of characters.

Then, model training is done using the Trainer API provided by the Hugging Face library. During the training phase, hyperparameters like batch size, learning rate, and number of epochs are adjusted according to the requirements of the model. Once trained, the model predicts whether the text falls under one of the following categories: Normal, Depression, or Suicide Risk. After training, the model and tokenizer are saved to be used in future real-time predictions on unseen texts.

C. Audio Processing Module

The design of this module aims to process the input data in an efficient manner, regardless of whether the data is in an uploaded file form or has been recorded in real-time. In the initial phase of pre-processing, the input data is prepared in the form of audio and processed through a cleaning phase that involves stripping off unnecessary noise as well as removing redundant portions of the input. The next step involves extracting relevant features using MFCCs, after which they will be processed by the machine learning model.

D. System Integration and Interface

It was developed with the use of Streamlit, which ensures a well-structured user interface. Multiple tabs have been used in its design, helping users to easily carry out text and voice analyses while also visualizing them in the form of graphs. Different types of output data are available through the application, including class predictions, confidence scores, and risk scores. This information is complemented with gauges and color schemes to make the result more interpretable.

E. Performance Optimization

For the smooth operation of the system, it has been optimized to function effectively in environments with CPUs. This way, no special hardware is needed for this purpose. Proper attention has been paid to balancing model precision and processing time based on the amount of data used for training. Additionally, the tool has been optimized for fast inference.

V. RESULTS AND DISCUSSION

The effectiveness of the proposed approach is evaluated using key performance metrics such as accuracy, precision, recall, and F1-score. In this analysis, the best-performing results obtained from both the BERT-based text model and the audio emotion recognition model across all experiments are taken into consideration.

Table 3: Key performance metrics of BERT text classification model

Class	Precision	Recall	F1-Score
Normal	0.95	0.94	0.94
Depression	0.73	0.62	0.67
Suicide Risk	0.94	0.96	0.95
Overall Accuracy			0.94

These outcomes prove that the system functions consistently in terms of accuracy and speed. In general, the models display a high level of competence in converting gestures to text and speech.

Table 4: Summary performance of audio emotion recognition model

Metric	Value
Accuracy	0.95
Macro F1-Score	0.95
Weighted F1-Score	0.95

The audio model achieves 95% accuracy, indicating reliable detection of emotional states from speech. This strengthens the system by providing additional context beyond text.

VI. CONCLUSION AND FUTURE ENHANCEMENTS

The proposed model is an ensemble of two methods that work in synergy: text classification through pre-training with BERT and emotion recognition based on speech signals. Combining the analysis of linguistic information with the analysis of vocal features allows us to obtain very good results the accuracy rate of text classification is around 94%, while emotion recognition for speech is approximately 95%.

However, there are some issues that have been detected during testing of the model. It is difficult for the machine learning model to classify depression correctly since this state is expressed implicitly and can be indicated by a lot of nuances in text or speech.

Another problem of this method is related to sentences with complex negations or several different emotions in the same sentence. There are several ways to improve the machine learning model in order to ensure better efficiency. Firstly, it is crucial to increase the diversity of the sample. Secondly, one should implement explainable AI methods in order to make the algorithm more transparent. Additionally, one should improve the interpretation of complex language structures like negation and mixed emotions.

REFERENCES

[1] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying Mental Health Signals in Twitter," in *Proc. ACL Workshop*, 2015.

[2] A. Yates, A. Cohan, and N. Goharian, "Depression and Self-Harm Risk Assessment in Online Forums," in *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[3] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

[4] M. Matero, A. Idnani, Y. Son, *et al*., "Suicide Risk Assessment with Multi-Level Dual-Context Language and BERT," in *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[5] F. Eyben, K. Scherer, B. Schuller, *et al*., "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS)," *IEEE Transactions*, 2016.

[6] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of Depression-Related Posts in Reddit Using Deep Learning," *IEEE Access*, 2019.

[7] A. H. Orabi, P. Buddhitha, M. H. Orabi, and D. Inkpen, "Deep Learning for Depression Detection in Twitter," in *Proc. CLPsych Workshop*, 2018.

[8] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, "Detecting Depression and Mental Illness on Social Media," *Current Opinion in Behavioral Sciences*, 2017.

[9] T. Al Hanai, M. Ghassemi, and J. Glass, "Detecting Depression with Audio/Text Sequence Modeling," in *Proc. Interspeech*, 2018.

[10] R. Sawhney, P. Manchanda, and A. Aggarwal, "Multimodal Depression Detection Using Deep Learning," *IEEE*, 2021.

[11] M. A. Bhat and M. M. Siddi, "Real-Time Kannada Sign Language Recognition to Speech and Text for Rural Primary Healthcare Centers," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 11, no. 7, pp. 35–40, Jul. 2022.