© JAN 2026 | IRE Journals | Volume 9 Issue 7 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I7-1713848

An Emotion-Aware AI-Based Virtual Therapy System Using Text and Speech Analysis

ANUSHKA BOHRA
Department of Computer Science and Engineering, KCC Institute of Technology and Management, Greater Noida, Uttar Pradesh, India

Abstract- ***Mental health support systems based on conventional chatbots often fail to provide emotionally appropriate responses due to their limited understanding of user emotions. To address these limitations, this paper presents an emotion-aware virtual therapy system that utilizes both textual and speech input to detect the emotional state of users and generate context sensitive therapeutic responses. The proposed system integrates natural language processing techniques for text-based emotion recognition and deep learning models for speech emotion analysis using acoustic features such as melfrequency cepstral coefficients. A multimodal architecture is designed to improve overall accuracy and reliability. The system is implemented as a mobile application to ensure accessibility and real time interaction. Experimental evaluation conducted on benchmark emotion datasets demonstrates that the proposed approach achieves higher performance compared to unimodal emotion detection methods. The results indicate that incorporating emotion awareness significantly enhances the effectiveness of virtual therapy systems by enabling personalized and empathetic user interactions. This work highlights the potential of multimodal artificial intelligence in scalable and intelligent mental health support solutions.***

Keywords— Emotion Detection, Mental Health Support, Virtual Therapy System, Natural Language Processing, Speech Emotion Recognition, Multimodal Deep Learning, AI Chatbot

I. INTRODUCTION

Mental health disorders have become one of the most significant global challenges, affecting millions of individuals across different age groups and socio-economic backgrounds. The increasing demand for psychological support has exposed the limitations of traditional therapy methods, including high costs, limited availability of trained professionals, and social stigma associated with seeking help. As a result, digital mental health solutions such as virtual assistants and chatbot-based therapy systems have gained considerable attention in recent years. These systems aim to provide accessible and immediate support to users through conversational interfaces. However, most existing solutions lack emotional intelligence and fail to accurately understand the emotional state of users, which reduces the effectiveness of human–computer interaction in sensitive mental health contexts.

Understanding human emotions is a critical aspect of effective therapeutic communication. Users often express their psychological state through both text and speech, making emotion recognition a challenging multimodal problem. Many existing systems rely only on text-based sentiment analysis, which is insufficient to capture complex emotional cues such as stress, anxiety, or sadness reflected in voice signals. To overcome these limitations, this paper proposes an emotion-aware virtual therapy system that analyses both textual and speech inputs using natural language processing and deep learning techniques. The objective of this work is to design an intelligent system capable of detecting user emotions in real time and generating appropriate therapeutic responses, thereby improving user engagement, empathy, and overall effectiveness of AI-based mental health support.

II. RELATED WORK

A. Text-Based Emotion Recognition in Mental Health Systems

Several studies have explored the use of natural language processing techniques to detect emotions from user-generated text in mental health applications. Early approaches employed lexicon-based methods and traditional machine learning classifiers such as support vector machines and naïve Bayes to categorize emotional states [1]. With the emergence of deep learning, recurrent neural networks and long short-term memory models have been widely adopted to capture

sequential dependencies in conversations [2]. More recently, transformer-based architectures and pre-trained language models such as BERT have demonstrated significant improvements in contextual understanding and emotion classification accuracy [3]. However, these systems remain limited in scenarios where emotional cues are implicitly expressed or masked within short or ambiguous textual messages.

B. Speech Emotion Recognition Using Deep Learning

Speech emotion recognition has gained increasing attention due to its ability to capture paralinguistic features such as tone, pitch, and speaking rate, which are strong indicators of human emotions. Researchers commonly extract acoustic features including Mel-frequency cepstral coefficients, spectral contrast, and zero-crossing rate, followed by classification using convolutional neural networks or hybrid CNN–LSTM models [4]. These approaches have achieved promising results on benchmark datasets such as RAVDESS and IEMOCAP [5]. Nevertheless, most speech-based systems are designed as standalone emotion classifiers and are rarely integrated into complete therapeutic or conversational frameworks.

C. Multimodal Emotion-Aware Therapy Systems

To address the limitations of unimodal approaches, recent studies have proposed multimodal systems that combine textual and speech-based emotion recognition [6]. Such systems aim to improve robustness by leveraging complementary information from multiple modalities. Some emotion-aware chatbots have incorporated cognitive behavioral therapy strategies to enhance user engagement and psychological outcomes [7]. Despite these advances, challenges remain in terms of real-time processing, deployment on mobile platforms, and personalization of therapeutic responses based on long-term emotional patterns. Existing solutions often lack scalability and practical usability in real-world mental health support environments [8].

III. PROPOSED SYSTEM

The proposed emotion-aware virtual therapy system is designed to detect user emotions from both text and speech inputs and generate adaptive therapeutic responses accordingly. The system follows a modular architecture consisting of data acquisition, emotion recognition, response generation, and user interaction layers. A multimodal approach is adopted to improve robustness and accuracy by combining complementary information from textual and acoustic signals. The overall workflow of the system is illustrated in Fig. 1.

A. System Architecture

The system architecture is composed of four main layers: the user interface layer, preprocessing layer, emotion detection layer, and therapy response layer. Users interact with the system through a mobile application interface that supports both text and voice-based communications. Textual input is directly forwarded to the natural language processing module, while speech input is first processed by a speech-to-text engine and an audio feature extraction module.

The preprocessing layer cleans and normalizes the input data by removing noise, stop words and irrelevant symbols for text and by filtering background noise for speech signals. The emotion detection layer employs deep learning models to classify emotions such as happiness, sadness, anger and fear. Finally, the therapy response layer selects and generates suitable responses based on the detected emotional state and conversation context.

B. Text-Based Emotion Recognition Module

The text-based emotion recognition module analyzes user messages using natural language processing techniques. Input sentences are tokenized, normalized and transformed into numerical representations using word embeddings or contextual embeddings obtained from pre-trained language models. A deep learning classifier is then applied to predict the underlying emotional category of the text.

This module enables the system to understand semantic and contextual cues present in user messages, such as expressions of stress, anxiety, or frustration. The predicted emotion label is forwarded to the multimodal fusion module for final decision making.

© JAN 2026 | IRE Journals | Volume 9 Issue 7 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I7-1713848

C. Speech-Based Emotion Recognition Module

The speech-based emotion recognition module processes voice inputs to extract acoustic features that reflect emotional characteristics. Features such as Mel-frequency cepstral coefficients are computed from the speech signal and provided as input to a convolutional neural network or a hybrid CNN–LSTM model.

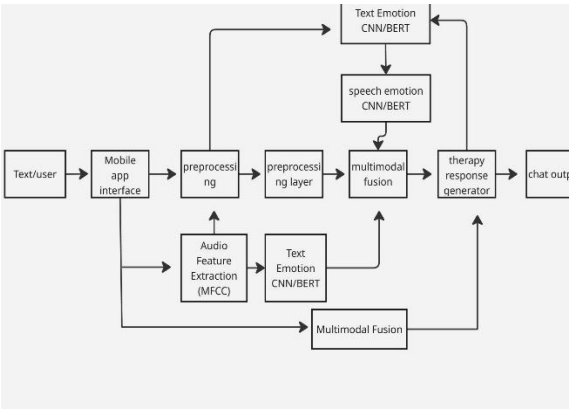

Fig. 1. Architecture of the proposed emotion-aware virtual therapy system.

IV. METHODOLOGY AND EXPERIMENTAL SETUP

This section describes the datasets used, data preprocessing steps, feature extraction techniques, model architecture, training procedure, and evaluation metrics employed to validate the proposed emotion-aware virtual therapy system.

A. Datasets

To evaluate the performance of the proposed system, publicly available benchmark datasets were used for both text-based and speech-based emotion recognition.

For text emotion recognition, the GoEmotions [9] dataset was utilized. This dataset consists of thousands of English sentences annotated with multiple emotion categories such as happiness, sadness, anger, fear, surprise, and neutrality. It provides high-quality labeled data suitable for training deep learning models in conversational emotion detection tasks.

For speech emotion recognition, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [10] dataset was used. It contains audio recordings from multiple speakers expressing different emotions including calm, happy, sad, angry, fearful, and neutral. The dataset is widely used in emotion recognition research and provides balanced and noise-controlled samples.

Table1 *Description of datasets used:*

Dataset Name	Modality	No. of samples	Emotion classes	Purpose
GoEmotions	Text	58,000+ sentences	6(happy,sad, anger, fear, surprise,neutral)	Text emotion
RAVDESS	Speech	1440 audio files	6	Speech emotion

B. Data Preprocessing

Text data preprocessing involved converting all characters to lowercase, removing punctuation and special symbols, tokenization and elimination of stop words. The cleaned text was then converted into numerical representations using word embeddings or contextual embeddings.

For speech data, audio signals were resampled to a uniform sampling rate and background noise was reduced using spectral filtering techniques. Each audio file was segmented into fixed-length frames to ensure uniformity across samples.

C. Feature Extraction

For textual input, semantic features were extracted using embedding techniques such as Word2Vec or transformer-based embeddings obtained from pre-trained language models.

For speech input, Mel-Frequency Cepstral Coefficients (MFCCs) were extracted as primary acoustic features. MFCCs effectively represent the short-term power spectrum of speech and are widely used in speech emotion recognition systems.

Table 2 Extraction Features for Emotion Recognition

Modality	Feature Type	Description
Text	Word embeddings/ BERT embeddings	Contextual semantic representation
Speech	MFCC	Represents short-term speech spectrum
Speech	Pitch	Frequency variation
Speech	Energy	Signal intensity

D. Model Architecture and Training

The text emotion recognition model employs a deep learning classifier based on either a Long Short-Term Memory network or a transformer-based architecture. The model processes embedded text sequences and outputs probabilities for each emotion class.

The speech emotion recognition model is implemented using a convolutional neural network combined with an LSTM layer to capture both spatial and temporal features from MFCC representations.

Both models were trained using categorical cross-entropy loss and optimized using the Adam optimizer. The datasets were divided into training, validation, and testing sets in an 80:10:10 ratio to ensure unbiased evaluation.

Table 3 *Model Architecture Parameters*

Parameters	Text Model	Speech Model
Model type	LSTM/ BERT	CNN + LSTM
Input size	Variable Length	40 MFCC coefficients
Hidden units	128	128
Optimizer	Adam	Adam
Loss Function	Categorical cross entropy	Categorical cross entropy
Epochs	20	25
Batch Size	32	32

E. Data Split Table

Table 4

Dataset	Training	Validation	Testing
GoEmotions	80%	10%	10%
RAVDESS	80%	10%	10%

V. CONCLUSION

This paper presented the design and implementation of an emotion-aware virtual therapy system that leverages both textual and speech inputs to identify user emotions and generate adaptive therapeutic responses. By integrating natural language processing techniques with deep learning-based speech emotion recognition models, the proposed system effectively captures both semantic and paralinguistic emotional cues. The multimodal fusion strategy further enhances robustness and accuracy compared to unimodal approaches.

Experimental evaluation on benchmark datasets demonstrated that the proposed framework achieves reliable performance in classifying emotional states and enables more empathetic and personalized interactions in mental health support applications. The mobile-based deployment of the system also improves accessibility and usability for users seeking immediate psychological assistance.

In future work, the system can be extended to support multilingual emotion recognition, continuous learning from user feedback, and integration with professional therapy guidelines. Additionally, incorporating physiological signals and long-term user behavior analysis may further enhance the accuracy and effectiveness of emotion-aware virtual therapy systems.

VI. ACKNOWLEDGMENT

The author would like to express sincere gratitude to the faculty members of the Department of Computer Science and Engineering, KCC Institute of

Technology and Management, Greater Noida, for their valuable guidance, continuous support, and constructive feedback throughout the course of this research work. Special thanks are extended to all peers and reviewers whose suggestions helped in improving the quality of this paper. The author also acknowledges the developers and researchers who made the publicly available datasets and tools used in this study possible.

REFERENCES

[1] B. Liu, "Sentiment analysis and opinion mining," Morgan & Claypool Publishers, 2012.

[2] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," EMNLP, 2015.

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," NAACL-HLT, 2019.

[4] Z. Zhao, Z. Bao, Z. Zhang, and N. Cummins, "Automatic speech emotion recognition using deep neural networks," IEEE ICASSP, 2014.

[5] S. Livingstone and F. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," PLoS ONE, 2018.

[6] S. Poria, E. Cambria, D. Hazarika, and N. Majumder, "Multimodal sentiment analysis: Addressing key issues and setting up baselines," IEEE Intelligent Systems, 2018.

[7] M. Inkster, S. Sarda, and V. Subramanian, "An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being," JMIR mHealth and uHealth, 2018.

[8] A. Abd-Alrazaq et al., "Effectiveness and safety of using chatbots to improve mental health: systematic review," Journal of Medical Internet Research, 2020.

[9] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A dataset of fine-grained emotions," in Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 4040–4054.

[10] S. Livingstone and F. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," PLoS ONE, vol. 13, no. 5, 2018.