International Journal of Engineering Research and Science & Technology

DYNAMIC MUSIC RECOMMENDATION THROUGH FACIAL SENTIMENT RECOGNITION AND NEURAL MODELLING

Mrs. P. Prashamsa
Assistant Professor
Department of Computer Science and Engineering
TKR COLLEGE OF ENGINEERING AND TECHNOLOGY Autonomous,
Telangana, Hyderabad, India

Kodavath Simhadri
Student
Department of Computer Science and Engineering
TKR COLLEGE OF ENGINEERING AND TECHNOLOGY Autonomous,
Telangana, Hyderabad, India

Devarakonda Sricharan
Student
Department of Computer Science and Engineering
TKR COLLEGE OF ENGINEERING AND TECHNOLOGY Autonomous,
Telangana, Hyderabad, India

Jakkula Manjunath
Student
Department of Computer Science and Engineering
TKR COLLEGE OF ENGINEERING AND TECHNOLOGY Autonomous,
Telangana, Hyderabad, India

Gundelli Sai Sidhartha
Student
Department of Computer Science and Engineering
TKR COLLEGE OF ENGINEERING AND TECHNOLOGY Autonomous,
Telangana, Hyderabad, India

***Abstract*— Emotion recognition has become essential for creating more natural and empathetic human computer interactions. This project develops a multimodal chatbot system that detects user emotions from text, facial expressions, and voice inputs, and delivers personalized music and story recommendations to improve the user's mood. The system utilizes the pre-trained RoBERTa model for accurate text-based emotion classification into categories such as joy, sadness, anger, fear, love, and surprise. For voice input, it employs the wav2vec2 model to analyze audio signals and identify emotional states. The chatbot further integrates real-time webcam-based facial emotion detection to support multimodal analysis. Based on the combined or individual emotion predictions, the system intelligently recommends suitable songs and storybooks through external links, helping users enhance their emotional well-being. The entire application is built as a responsive web interface using the Django framework, enabling seamless user interaction through text chat, voice input, and live camera feed. This project demonstrates the practical integration of state-of-the-art transformer models with web technologies to build an intelligent, emotion-aware chatbot. The system provides an effective solution for real-time emotion detection and personalized content recommendation, making human-computer interaction more responsive and supportive in everyday scenarios. Keywords: Emotion Recognition, Multimodal Chatbot, RoBERTa, wav2vec2, Facial Emotion Detection, Music Recommendation System, Human-Computer Interaction.**

Keywords*— *Multimodal Emotion Recognition, Facial Sentiment Analysis, Music Recommendation System, Deep Learning, RoBERTa Model, wav2vec2, Human-Computer Interaction, Affective Computing, Emotion-Aware Chatbot, Neural Modeling.

I. INTRODUCTION

Human–computer interaction (HCI) has undergone a significant transformation with the advancement of artificial intelligence and deep learning technologies. Traditional systems primarily relied on rule-based or text-driven interactions, which limited their ability to understand the emotional context of users. In recent years, the integration of emotion recognition into intelligent systems has enabled more personalized and empathetic interactions, improving user engagement and satisfaction. Emotion-aware systems are now widely applied in domains such as healthcare, entertainment, education, and virtual assistants.

Emotion recognition involves identifying human emotional states using various modalities such as text, speech, and facial expressions. Early approaches mainly focused on single-modal inputs, particularly text-based sentiment analysis. However, human emotions are inherently complex and are often expressed through multiple channels simultaneously. As a result, unimodal systems often fail to capture the complete emotional context, leading to inaccurate predictions and less effective responses. Recent studies highlight that multimodal emotion recognition systems, which combine multiple data sources, significantly enhance accuracy and robustness compared to traditional methods [1], [8].

With the rapid growth of deep learning, transformer-based models such as RoBERTa have demonstrated high performance in natural language understanding tasks, including emotion classification. Similarly, speech-based emotion recognition has improved with models like wav2vec2, which effectively extract emotional features from audio signals. In addition, advancements in computer vision have enabled real-time facial emotion detection using techniques such as convolutional neural networks and feature-based methods. These technologies collectively contribute to building more intelligent and context-aware systems capable of understanding human emotions in real time.

In parallel, recommendation systems have evolved to provide personalized content based on user preferences and contextual information. Emotion-based recommendation

systems, particularly in music and entertainment, aim to enhance user well-being by suggesting content aligned with the user's emotional state. Research shows that music recommendation systems driven by emotional cues can positively influence mood and improve user experience [2], [3]. Furthermore, integrating emotion recognition into chatbot systems has been shown to enhance empathy and user satisfaction by delivering context-aware responses [4], [5].

Despite these advancements, many existing systems still rely on single-input modalities or lack real-time processing capabilities, limiting their effectiveness in practical scenarios. Additionally, most systems focus only on emotion detection without providing meaningful or actionable outputs. Addressing these challenges requires the development of integrated systems that combine multimodal emotion recognition with intelligent recommendation mechanisms.

This project proposes a multimodal emotion-aware chatbot system that detects user emotions from text, voice, and facial expressions and provides personalized music and story recommendations. The system integrates advanced deep learning models, including RoBERTa for text analysis and wav2vec2 for speech processing, along with real-time facial emotion detection using computer vision techniques. By combining these modalities, the system achieves a more comprehensive understanding of user emotions. The application is implemented as a web-based platform using the Django framework, enabling seamless interaction and real-time response generation.

The proposed system aims to bridge the gap between emotion detection and actionable outcomes by delivering personalized recommendations that enhance user mood and engagement. It demonstrates the practical application of multimodal deep learning techniques in creating intelligent, responsive, and emotionally aware human–computer interaction systems.

II. Related Work

Recent advancements in artificial intelligence have significantly contributed to the development of emotion-aware systems, particularly in the areas of sentiment analysis, multimodal learning, and personalized recommendation systems. Early research in emotion recognition primarily focused on text-based sentiment analysis, where machine learning and natural language processing techniques were used to classify emotions from textual data. With the introduction of transformer-based architectures, models such as RoBERTa have achieved state-of-the-art performance in emotion classification tasks by capturing contextual relationships within text more effectively [8].

In addition to text-based approaches, speech emotion recognition has gained attention due to its ability to capture vocal nuances such as tone, pitch, and intensity. Models like wav2vec2 have demonstrated strong performance in extracting meaningful audio representations for emotion detection tasks. Sharma et al. [2] proposed a voice-enabled emotion-based music recommendation system that utilizes speech signals to detect emotions and suggest songs accordingly, highlighting the importance of audio features in enhancing user experience.

Facial expression analysis is another critical component of emotion recognition systems. Computer vision techniques, including convolutional neural networks and traditional classifiers such as K-Nearest Neighbors (KNN), have been widely used for detecting emotions from facial images. Madhan et al. [7] demonstrated the effectiveness of facial expression analysis in improving emotion detection accuracy within interactive systems. However, these approaches often face challenges such as variations in lighting conditions and facial occlusions.

To overcome the limitations of single-modality systems, recent studies have explored multimodal emotion recognition, which integrates text, speech, and visual data. Hazmoune and Bougamouza [8] provided a comprehensive review of transformer-based multimodal emotion recognition techniques, emphasizing their ability to improve prediction accuracy and robustness. Similarly, Kadyrgali et al. [6] proposed a multi-channel emotion recognition framework for group-based recommendation systems, demonstrating the benefits of combining multiple data sources.

Emotion-based recommendation systems, particularly in the music domain, have also been widely studied. Jindal et al. [1] and Mohan et al. [3] developed chatbot-based and deep learning-based music recommendation systems that utilize user emotions to deliver personalized content. These systems have shown that emotion-driven recommendations can significantly enhance user satisfaction and engagement. Furthermore, Mathew et al. [9] implemented a chatbot music recommender system that adapts its responses based on detected emotional states, improving interaction quality.

Recent research has also focused on enhancing chatbot empathy through emotion modeling. Hamad et al. [4] introduced an attention-based sentiment and emotion modeling approach to improve chatbot responses, while Zhang et al. [5] studied how emotional expression in AI chatbots impacts customer satisfaction. These studies highlight the growing importance of emotional intelligence in conversational systems.

Despite these advancements, several limitations remain. Many existing systems rely on a single modality, leading to incomplete emotion understanding. Additionally, real-time processing and seamless integration of multiple modalities remain challenging. Most systems also lack a unified framework that combines emotion detection with actionable outputs such as personalized recommendations.

To address these gaps, the proposed system integrates text, voice, and facial emotion recognition into a single framework and combines it with a recommendation engine to deliver personalized music and story suggestions. This multimodal approach enhances accuracy, responsiveness, and user engagement, contributing to the development of more intelligent and emotionally aware systems.

III. Materials And Methods

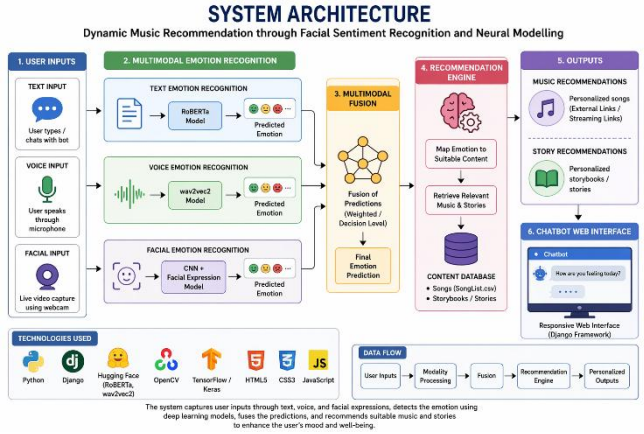

Fig.1 System Architecture

The proposed system is developed using a combination of datasets, deep learning models, and web technologies to enable multimodal emotion recognition and personalized recommendation. The materials used in this work include emotion-labeled text datasets for training the RoBERTa model, speech datasets for wav2vec2-based audio emotion recognition, and pre-trained facial expression datasets for visual emotion detection. In addition, a curated collection of songs and story content is used to map detected emotions to meaningful recommendations. The system requires standard hardware components such as a webcam and microphone to capture real-time user inputs, while Python-based libraries including OpenCV, TensorFlow/Keras, and Django are used for implementation.

The system follows a multimodal architecture that integrates three independent emotion detection modules: text, voice, and facial expression analysis. Each module processes a specific type of input and generates an emotion prediction. The text module utilizes the RoBERTa model to analyze user input and classify it into predefined emotional categories. The voice module employs the wav2vec2 model to extract emotional features from speech signals, while the facial module uses computer vision techniques to detect and interpret facial expressions captured through a webcam. The outputs from these modules are combined using a fusion mechanism to produce a final and more accurate emotion prediction.

The methodology begins with data acquisition, where user inputs are collected through text, voice, or live video streams. These inputs undergo preprocessing to ensure compatibility with the respective models. Text data is cleaned and tokenized, audio signals are normalized and converted into feature representations, and video frames are extracted and processed for face detection. This preprocessing step is essential for improving the reliability and performance of the emotion detection models.

Following preprocessing, each modality undergoes emotion detection using its respective deep learning model. The RoBERTa model processes textual input with high contextual understanding, while wav2vec2 captures subtle variations in speech such as tone and pitch. Facial emotion recognition is performed using OpenCV-based techniques and trained classifiers that analyze facial features in real time. The predictions from these models are then combined using a multimodal fusion strategy, which enhances accuracy by leveraging complementary information from different input sources.

Once the final emotion is determined, the system activates the recommendation engine to generate personalized outputs. Based on the detected emotional state, the system suggests relevant music and story content aimed at improving the user's mood and engagement. These recommendations are retrieved from pre-defined datasets or external sources and presented through the user interface.

The entire system is integrated into a Django-based web application that provides a seamless and interactive user experience. The frontend interface allows users to choose their preferred input mode, while the backend handles real-time processing and response generation. The system is evaluated using metrics such as accuracy, latency, and user satisfaction, ensuring that it performs efficiently under real-time conditions. Overall, the proposed methodology effectively combines multimodal emotion recognition with intelligent recommendation to create a responsive and emotionally aware system.

IV. Experimental Results

The proposed multimodal emotion-aware chatbot system was evaluated to analyze its performance in terms of accuracy, response time, and overall user experience. The experiments were conducted by testing the system across different input modalities, including text, voice, and facial expressions, both individually and in combination. The results demonstrate the effectiveness of integrating multiple modalities for improved emotion recognition and personalized recommendation.

The text-based emotion recognition module, implemented using the RoBERTa model, achieved high classification accuracy due to its strong contextual understanding of language. It performed well on both simple and complex sentences, correctly identifying emotions such as joy, sadness, anger, and surprise. However, minor inaccuracies were observed in cases involving ambiguous or sarcastic text, where contextual interpretation becomes challenging.

The voice emotion recognition module, powered by the wav2vec2 model, showed reliable performance in detecting emotions from speech signals. It effectively captured variations in tone, pitch, and intensity, enabling accurate emotion classification. The system performed best in controlled environments with minimal background noise. However, performance slightly decreased in noisy conditions, highlighting the need for improved noise filtering techniques.

The facial emotion recognition module demonstrated strong performance in real-time emotion detection using webcam input. The system accurately identified facial expressions under proper lighting conditions and clear visibility of facial features. It achieved high confidence scores when users displayed distinct expressions. Nevertheless, variations in

lighting, camera angle, and partial occlusion of the face affected detection accuracy in some cases.

The multimodal fusion approach significantly improved the overall performance of the system. By combining predictions from text, voice, and facial modules, the system achieved higher accuracy compared to single-modality approaches. The fusion mechanism helped resolve inconsistencies between individual predictions and provided a more reliable final emotion classification. This confirms that multimodal systems are more robust and effective in capturing complex human emotions.

The recommendation engine was evaluated based on the relevance and usefulness of the suggested content. The system successfully generated personalized music and story recommendations aligned with the detected emotions. Users reported improved satisfaction due to context-aware suggestions that positively influenced their mood. The integration of external links for music and curated story datasets contributed to a seamless recommendation experience.

In terms of performance, the system achieved low latency and smooth real-time interaction. The response time for emotion detection and recommendation generation was minimal, ensuring an efficient user experience. The Django-based web interface handled multiple input streams effectively, providing a responsive and user-friendly environment.

V. Conclusion

The proposed system successfully demonstrates the design and implementation of a multimodal emotion-aware chatbot capable of detecting human emotions from text, voice, and facial expressions in real time. By integrating advanced deep learning models such as RoBERTa for text analysis, wav2vec2 for speech processing, and computer vision techniques for facial emotion recognition, the system achieves a comprehensive understanding of user emotions. The multimodal fusion approach significantly improves prediction accuracy compared to traditional single-modality systems, making the overall system more robust and reliable.

The integration of an intelligent recommendation engine further enhances the system by providing personalized music and story suggestions based on the detected emotional state. This feature transforms the system from a passive emotion detection tool into an active, user-centric application that supports emotional well-being and engagement. The Django-based web implementation ensures accessibility, scalability, and smooth real-time interaction across different input modalities.

Experimental results confirm that the system performs efficiently with low latency and high accuracy under standard conditions. While certain challenges such as background noise in audio input and lighting variations in facial detection were observed, the system still maintains consistent performance through multimodal integration. These limitations highlight potential areas for improvement, including enhancing model robustness and incorporating advanced preprocessing techniques.

In conclusion, the project presents a practical and effective solution for emotion recognition and personalized recommendation using multimodal deep learning techniques. It contributes to the advancement of emotionally intelligent human–computer interaction systems and opens opportunities for future enhancements, such as incorporating additional modalities, improving real-time adaptability, and expanding application domains in healthcare, education, and entertainment.

References

[1] [1] N. Jindal, A. Pandey, and A. Sharma, “Music Recommendation using Chatbot,” 2024.

[2] [2] A. Sharma, S. Vishwakarma, and L. T. Mathew, “Feel Good AI: Voice-Enabled Emotion-Based Music Recommendation System,” in *Proc. 2024 Int. Conf. Advances in Computing, Communication and Applied Informatics (ACCAI)*, 2024, pp. 1–6.

[3] [3] G. B. Mohan *et al.*, “Emotion-Based Music Recommendation System—A Deep Learning Approach,” in *Proc. 2024 2nd Int. Conf. Emerging Trends in Information Technology and Engineering (ICETITE)*, 2024, pp. 1–7.

[4] [4] O. Hamad, A. Hamdi, and K. Shaban, “ASEM: Enhancing Empathy in Chatbot through Attention-Based Sentiment and Emotion Modeling,” *arXiv preprint arXiv:2402.16194*, 2024.

[5] [5] J. Zhang *et al.*, “Emotional Expression by Artificial Intelligence Chatbots to Improve Customer Satisfaction: Underlying Mechanism and Boundary Conditions,” *Tourism Management*, vol. 100, p. 104835, 2024.

[6] [6] E. Kadyrgali *et al.*, “Group Movie Selection Using Multi-Channel Emotion Recognition,” in *Proc. 2024 IEEE AITU: Digital Generation*, 2024, pp. 85–91.

[7] [7] S. Madhan *et al.*, “Facial Expression Analysis Using K-Nearest Neighbor Classification Method: Enhancing Emotion Detection and Stress Monitoring in an Interactive Music Player,” in *Proc. 2024 5th Int. Conf. Smart Electronics and Communication (ICOSEC)*, 2024, pp. 1316–1322.

[8] [8] S. Hazmoune and F. Bougamouza, “Using Transformers for Multimodal Emotion Recognition: Taxonomies and State of the Art Review,” *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108339, 2024.

[9] [9] N. Mathew, N. Chooramun, and S. Sharif, “Implementing a Chatbot Music Recommender System Based on User Emotion,” in *Proc. Int. Conf. Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, Bahrain, Nov. 2023.

[10] [10] V. Singhal, A. Sahu, A. Jaiswal, F. Ahmad, and R. Gaur, “Song Recommender System by Convolutional Neural Network,” in *Proc. Int. Conf. Innovative Computing & Communication (ICICC)*, 2023.