Industrial Engineering Journal

ISSN: 0970-2555

Volume : 54, Issue 2, February : 2025

PERSONALIZED MUSIC RECOMMENDATION THROUGH MULTI-MODAL EMOTION RECOGNITION

Mrs Y V N Tulasi[1], Matta Neeraja[2], Kovilakaru Neha[3], Madala Venkata Adithya[4], Katragadda Harshith Ram[5] 1- Assistant Professor, 2,3,4,5- IV-B. Tech CSE Students Department of Computer Science and Engineering, Seshadri Rao Gudlavalleru Engineering College (An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada), Seshadri Rao Knowledge Village, Gudlavalleru521356, Andhra Pradesh, India.

***ABSTRACT*—We introduce, in this work, a holistic approach for music recommendation by multi-modal emotion recognition. With facial expressions, speech, and text sentiment analysis as the three inputs, our system dynamically matches the music recommendation with the emotional state of the user. This system ensures accuracy with real-time performance due to its use of the latest machine learning techniques, including CNNs, LSTMs, and pre-trained NLP models. Its effectiveness is well supported through experiments that compare it with conventional methods and provide a personalized and emotionally aware user experience.**

Keywords—Multi-modal emotion recognition, Personalized music recommendation, Facial expressions, Speech emotion analysis, Sentiment analysis.

I. INTRODUCTION

Technology advancement has been so fast that it has changed consumerism behaviour. People are being entertained differently and are involved in various activities considerably. Music is a pan-human music medium and is highly emotional. Hugely personalized, as it varies according to everyone's state of mind. However, most existing music recommendation systems can only apply collaborative filtering and content-based filtering. Although these systems can effectively trace patterns in user preferences, it is not enough to support the complexities regarding the interactions between music and human emotions.

The integration of emotion recognition into music recommendation systems provides a new sense of personalization: alignment of recommendations to users' current mood and feelings. Multi-modal emotion recognition is an emerging field that integrates facial expression analysis with speech emotion detection and text-based sentiment analysis using advancement in artificial intelligence. Since multi-modal systems are totally different from other single-modal systems, they offer a richer and more accurate assessment of emotions by taking into account information from a variety of sources.

This would be music for the soul; it is said that it determines and reflects the human mind, from sorrowful melodies which can give peace during grief times to a euphoric track that energizes and lifts the person up. The study recommended in this article will take up this emotional interaction by suggesting to the user what music to play according to his detected emotional states in real-time. In return, this increases the experience while listening and evokes an emotionally deeper relationship with the user to the music site.

This research presents a new system based on state-of-the-art technologies, such as CNNs, RNNs, LSTM models, and transformer-based NLP models. The system will integrate these technologies in order to analyze facial expressions, speech patterns, and text inputs, generating a complex emotional profile for each user. Such diversity in the system's data types makes it very adaptable to diverse user preferences and contexts.

The motivation for this study emanates from the increasing demand for hyper-personalized experiences in this digital age. With this research work, an effort is made to bridge the gap between technological innovation and emotional intelligence in the area of music recommendation systems. Not only does this research address technical challenges related to emotion detection and

recommendation mapping, but it also explores the very profound impact that such systems could have on user engagement and satisfaction.

This multi-modal emotion recognition framework is expected to be an accurate and scalable push beyond the boundaries of the conventional recommendation methodologies. By combining human-centric design with artificial intelligence, this system aims to build a seamless, emotionally enriching user experience while setting the scene for future innovation in the realm of personalized content delivery.

II. LITERATURE REVIEW

Emotion-based music recommendation systems attract much attention owing to their high potential to contribute to the richness of user experience by aligning music choices and emotional states. Traditional music recommendation systems relied primarily on collaborative filtering and content-based filtering approaches but recent advancements with deep learning has enabled the incorporation of multimodal data for personalized recommendations. This paper discussed an integrated facial expression, speech, and text-based approach to the enhancement of sentiment analysis for improved classification of emotion and showed how these multimodal approaches outperformed unimodal approaches for detection of users' emotions as well as recommending suitable items to users [1]. Facial emotion recognition has gained special attention and widespread researches over the last several years for their effectiveness in user mood detection, and for efficiency in facial features extraction, and classification through the use of convolutional neural networks. Deep learning-based frameworks have achieved significant improvements in accuracy and responsiveness in recommendation systems by mapping these detected emotions to music genres [2].

Speech emotion recognition has also emerged as a powerful tool in personalized music recommendations, leveraging recurrent neural networks (RNNs) and long short-term memory (LSTM) networks to extract emotional cues from vocal inputs. It has been established that addition of speech-based emotion recognition to a text sentiment analysis increases the robustness of the system, especially in scenarios where facial detection fails due to insufficient lighting or occlusions. Transformers and deep learning techniques have enabled more refined classification of emotions from speech tone and textual sentiment, leading to an improved user experience [3]. Real-time multimodal emotion recognition systems have also been developed. These systems aim to dynamically update recommendations based on facial expressions, voice tone, and text-based sentiment. It uses deep learning models to handle each modality separately before a decision fusion strategy is used for the integration of the results in order to remain flexible in terms of emotion detection and reduce the reliance on single input sources [4].

Recent research has been centered on context-aware personalized music recommendation, where deep learning models identify user behaviour, preferences, and contextual data to improve recommendations. In addition, hybrid facial, speech, and textual emotion recognition system has shown practicality in addressing issues such as user preference diversity and real-time adaptability. The approaches consequently enable music recommendation engines to present very accurate and emotionally relevant song selections, leading to an improved level of user engagement [5]. However, surveys on multimodal emotion-based recommendation systems show deep learning methods have gradually been adopted in the attention mechanism and self-supervised learning, which can enhance the accuracy of prediction. Future improvements are expected for optimization of these models through increased data and real-time processing capabilities for the best performance in emotion-based music recommendations [6].

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 2, February : 2025

III. METHODOLOGY

A. Data Collection

To guarantee that the system will be able to correctly identify emotions in different input modalities, the following datasets were used:

Facial Expression Recognition*:*

Dataset*:* FER2013 (Facial Expression Recognition 2013).

It Consists of 35,887 grayscale images (48x48 pixels) classified into seven emotions: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral.

Gathered from different real-world conditions, including various lighting, poses, and expressions.

Extensively used for training and testing emotion recognition models.

Speech Emotion Recognition:

Dataset: RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)

It comprises 1,440 high-quality emotional speech recordings performed by 24 actors (12 male, 12 female).

It covers seven emotions: Neutral, Happy, Sad, Angry, Fearful, Disgust, and Surprise.

It has high consistency in emotional delivery, which makes it a benchmark dataset for speech emotion recognition.

Text Emotion Recognition:

Pre-trained Model: Hugging Face DistilRoBERTa Fine-tuned on diverse text datasets such as GoEmotions and Crowdflower, which contain labeled text for emotions like Anger, Fear, Neutral, Sadness, Surprise, and Disgust.

Comprises text from social media, dialogues, and user-generated reports, which will ensure linguistic diversity and generalization.

B. Data Preprocessing

Each modality needed different preprocessing techniques to ensure consistency and compatibility with the emotion recognition models:

Facial Expression Preprocessing:

Resizing: Images were resized to 48x48 pixels to meet the input size requirement of the CNN model.

Grayscale Conversion: Images were normalized to grayscale to reduce computational complexity while retaining emotion-related features.

Normalization: Pixel intensity values were scaled to the range [0, 1].

Data Augmentation: Applied transformations like random rotations, flips, and brightness adjustments to improve model generalization.

Speech Emotion Preprocessing:

Feature Extraction: Extracted Mel-Frequency Cepstral Coefficients (MFCCs) from audio signals, capturing speech patterns relevant to emotions.

Noise Reduction: Applied audio filtering to reduce background noise and enhance clarity.

Normalization: MFCC features were standardized to zero mean and unit variance.

Segmentation: Long audio files were segmented into smaller chunks for better feature extraction and analysis.

Text Emotion Preprocessing:

Tokenization: Text input was tokenized using Hugging Face's tokenizer for DistilRoBERTa, where text was transformed into numerical sequences.

Padding and Truncation: Input sequences were padded or truncated to a fixed length of 512 tokens.

Cleaning: Removed special characters, URLs, and excessive whitespace for noise reduction.

C. Model Description

This system of multi-modal emotion recognition

leverages the specialization of architectures by input modality: CNN to detect facial emotion, LSTM for speech emotion recognition, and a transformer-based pre-trained model (DistilRoBERTa) for

sentiment analysis of the text. Each of these serves in tandem to establish an accurate perception of emotion-detection that eventually forms the premise for personalized recommendations of music.

1. Facial Emotion Recognition

Convolutional Neural Network (CNN)

The CNN model is a feature extractor that captures facial patterns and expressions from input images.

Architecture:

The CNN consists of multiple convolutional layers that extract spatial features, such as the curvature of a smile or furrowed brows, indicative of emotional states. Max-pooling layers reduce spatial dimensions, preserving critical features while minimizing computation.

Fully connected layers sum up the extracted features and project them to seven emotion classes (Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral).

Training Dataset:

The model is trained on the FER2013 dataset, which contains varied facial images under different real-world conditions.

2. Speech Emotion Recognition

Recurrent Neural Network (RNN) with LSTM

The LSTM-based RNN is used to extract sequential and temporal features from speech, capturing the emotional tone in audio signals.

Architecture:

Input: The features are the Mel-Frequency Cepstral Coefficients (MFCCs) extracted from audio signals.

LSTM Layers: On the sequential data of MFCCs, the LSTM layers move on to find temporal dependencies, like pitch and tone modulations that actually symbolize emotions.

Dense Layer: The out-module of LSTM is subsequently fed into a dense layer to classify audio into seven categories (Neutral, Happy, Sad, Angry, Fearful, Disgust, Surprise).

Softmax Layer : Outputs the probable for each class.

Training Data:

Trained on the RAVDESS dataset, comprising professionally recorded emotional speech.

3. Text-based Emotion Classification

Transformer-based Pre-trained Model (DistilRoBERTa)

The DistilRoBERTa model exploits the transformer architecture to process input text and pick up contextual meanings of emotions using words and phrases.

Architecture:

Pre-trained Transformer: DistilRoBERTa is a distilled version of RoBERTa, designed to be efficient during inference without a reduction in accuracy. This model uses self-attention that captures relationships among words in a sentence.

Tokenization: Input text is tokenized into numerical representations using a vocabulary-based tokenizer.

Embedding Layer: The model converts tokens into dense vector representations, preserving their semantic meaning.

Multi-Head Attention: The transformer applies multi-head attention to capture contextual dependencies within the text, such as sarcasm, subtle expressions, or mixed emotions.

Head for Classification: Text classification is output using a dense layer and the softmax activation function for seven different categories of emotion (Anger, Disgust, Fear, Joy, Neutral, Sadness, Surprise).

Pre-Trained Dataset:

GoEmotions dataset as well as others including Crowdflower, Emotion Dataset by Elvis et al.; datasets that used sources of varied texts such as Reddit, Twitter, and dialogue boxes.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 2, February : 2025

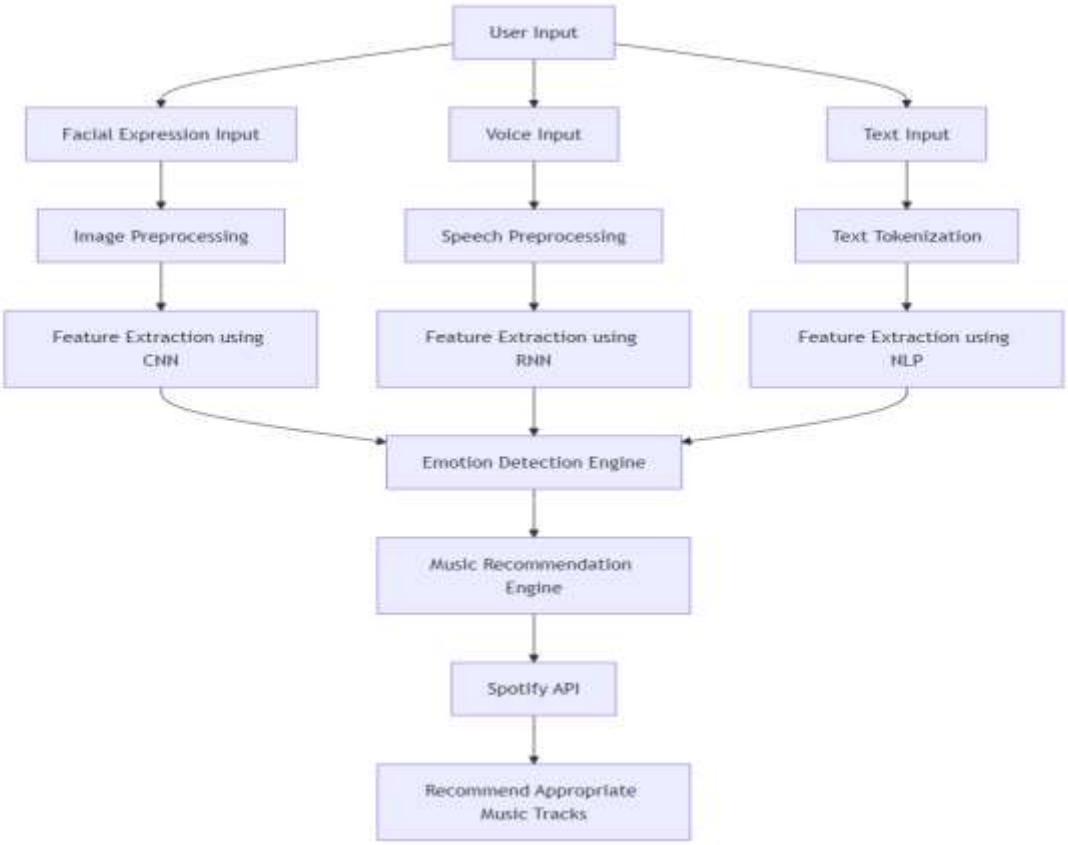


Fig1:- Flow diagram of Personalized Music Recommendation Through Multi-Modal Emotion Recognition

D. Implementation

The implementation entailed integrating the individual emotion detection models into one multi-modal emotion recognition and music recommendation system. The implementation steps involved:

1. Import Libraries

The following libraries and tools were used to build, train, and deploy the system:

TensorFlow and PyTorch: For model building and training the CNN, LSTM, and transformer models.
NumPy and Pandas: For numerical computations and data manipulation.
Matplotlib and Seaborn: for plotting the training performance and model evaluation.
Spotify: this is used to access the Spotify API to get the sound recommendations.
Flask: the back-end server and REST API that shall be used to receive user input and communicate with the models.
SQLAlchemy: the one used to authenticate users and manage the databases.
OpenCV and PIL: to handle and preprocess facial emotion input and images respectively.
SpeechRecognition and Librosa: for recording audio data, feature extraction, and preprocessing.
Hugging Face Transformers: To use the pre-trained DistilRoBERTa model for text emotion recognition.

2. Multi-Modal Input Handling:

A user-friendly interface was developed to allow users to select one of three input modes:
Facial Expressions: Captured using a webcam and processed in real time.
Speech: Recorded using a microphone and processed as audio files.
Text: Entered by the user in a text box.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 2, February : 2025

3. Emotion Detection Pipeline:

Depending on the modality selected by the user:

Facial Input: The CNN model takes the input as grayscale images for emotion prediction.

Speech Input: The LSTM model takes MFCC features of the audio file as input.

Text Input: The DistilRoBERTa model predicts the emotions of the tokenized text.

4. Emotion-to-Genre Mapping:

The detected emotions are mapped to pre-defined music genres through a genre mapping table.

Emotion	Genre(s)
Happy	Pop, Dance
Sad	Acoustic, Piano
Angry	Rock, Metal
Fearful	Ambient, Chill
Disgust	Alternative, Punk
Surprised	Indie, Electronic
Neutral	Classical, Jazz

5. Music Recommendation Engine:

Random Track Selection: Besides generating a list of tracks, the system also gives a random track suggestion from the queried results to enrich the user experience.

Tracks are presented to the user in a structured format.

6. Real-Time Processing:

Implemented asynchronous processing to handle real-time inputs from users:

Facial Expressions: The frames captured by the webcam are processed in real-time by the CNN model.

Speech: Audio files are buffered for a while, preprocessed, and fed into the LSTM model for instant processing.

Text: Typed input is tokenized and processed instantly using the Hugging Face DistilRoBERTa model.

7. Web Application Framework:

Developed the system using a Flask backend to integrate all functionalities, including emotion detection, music recommendation, and user interaction.

Designed a responsive frontend using HTML, CSS, and JavaScript, enabling users to interact seamlessly across devices.

Application pages:

Login and Registration Pages: Ensure secure user authentication.

Emotion Input Pages: Separate pages for face, speech, and text-based input interfaces.

Recommendation Results Page: Display the detected emotions and recommended tracks.

E. Model Fusion

The system combines the results of each of the individual models (face, speech, and text) into a single unified recommendation pipeline:

Input Selection:

One modality is selected by the user at any time such that the system only evaluates one type of input in any session.

Emotion Classification:

Each model independently classifies its input.

Music Query Generation:

The detected emotions are translated into genres of music, and a query is made to the Spotify API for pertinent tracks.

Output Display:

The results appear on the web interface, displaying the detected emotion, recommended songs, and a random track.

Industrial Engineering Journal

ISSN: 0970-2555

Volume : 54, Issue 2, February : 2025

IV. RESULTS

The system proposed here identifies emotions from three different sources: facial expressions, speech, and text. It integrates deep learning models and natural language processing techniques so that it can analyze user emotions and make recommendations for music appropriately.

Facial Emotion Prediction

- The model for facial emotion recognition achieved an accuracy of 83%.
- The trained deep learning model correctly classifies emotions into seven categories: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral.
- Preprocessed grayscale images (48x48) are used as input.
- The model is highly accurate in real-time predictions.

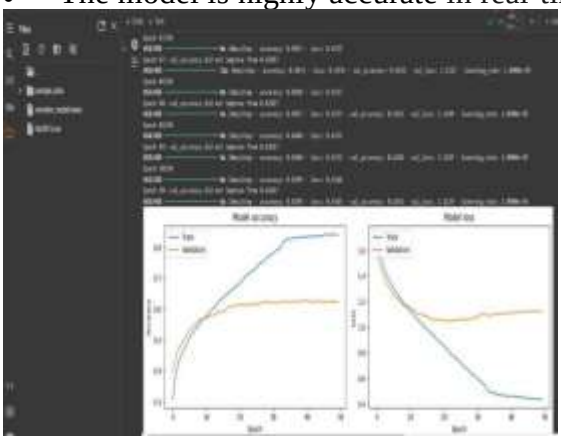


Fig: Accuracy values of CNN trained model for facial expression input.

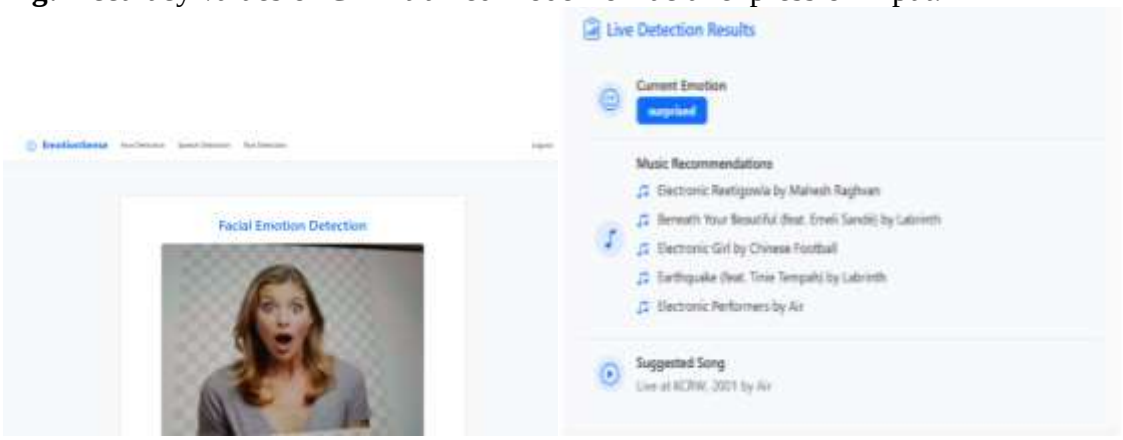


Fig: Facial Emotion Prediction Sample Output

Speech Emotion Prediction

- The model for speech emotion recognition achieved an accuracy of 90%.
- Speech recordings will be converted to text using Speech Recognition API.
- The text thus extracted will undergo sentiment and emotion classification.
- It has great accuracy in determining emotions from the speech.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 2, February : 2025

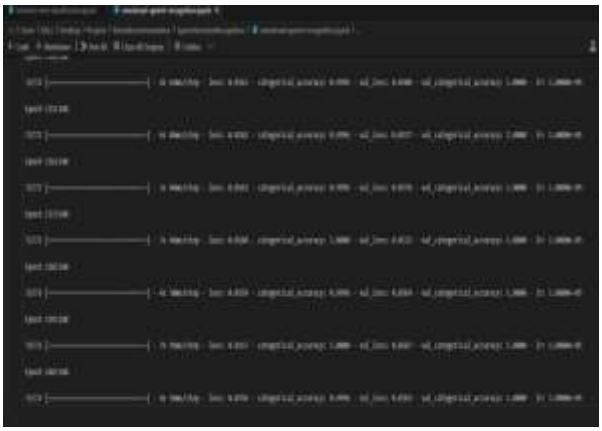


Fig: Accuracy values of RNN and LSTM trained models.

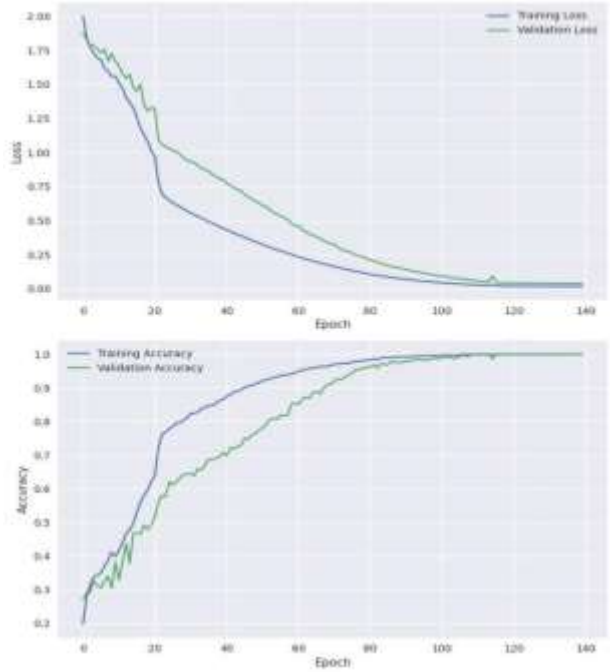


Fig: Model Training Accuracy and Loss Graphs

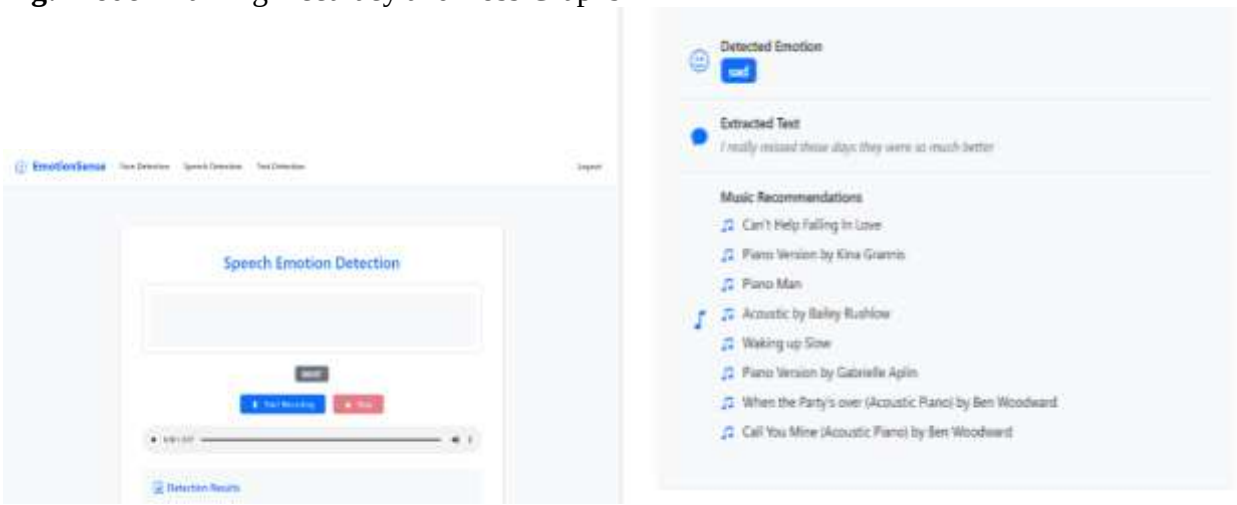


Fig: Speech Emotion Sample Output

- **Text Emotion Prediction**

The model for text emotion recognition achieved an accuracy of 90%.

- It uses DistilRoBERTa transformer-based model for the text emotion classification.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 2, February : 2025

- It includes the classification of emotions as follows: Anger, Disgust, Fear, Happy, Neutral, Sadness, and Surprise.

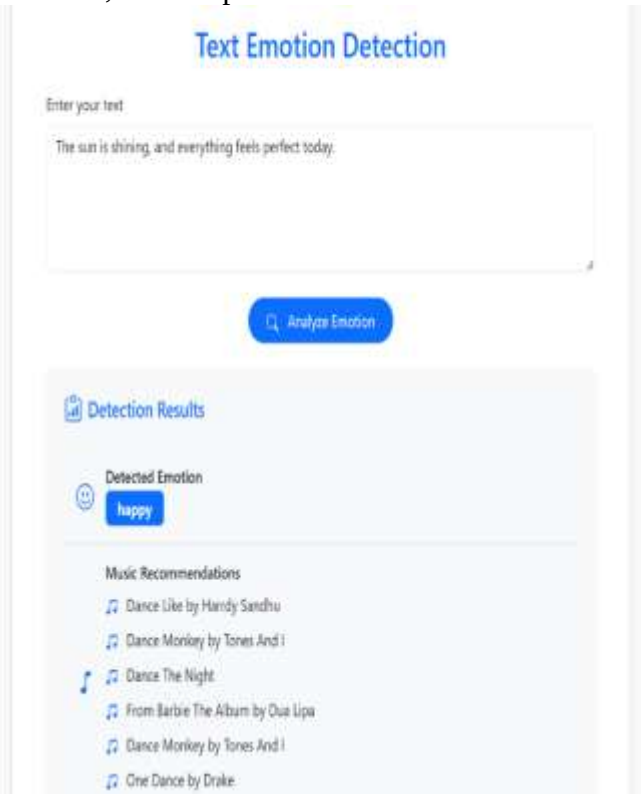


Fig: Recommend Music Tracks For Text Input.

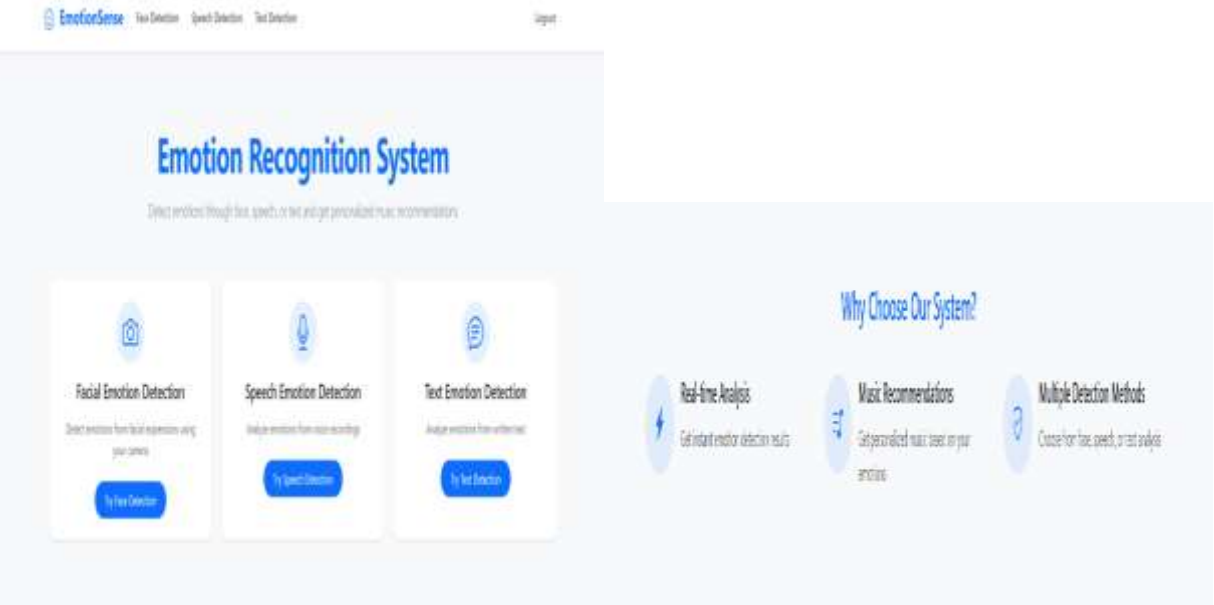
System Interface and User Interaction

The system offers a user-friendly interface in which users can choose between Facial, Speech, and Text-based emotion recognition. The interface consists of:

- Login Page for user authentication.
- Registration Page for new users to create an account.
- Home Page where users select the emotion recognition method.

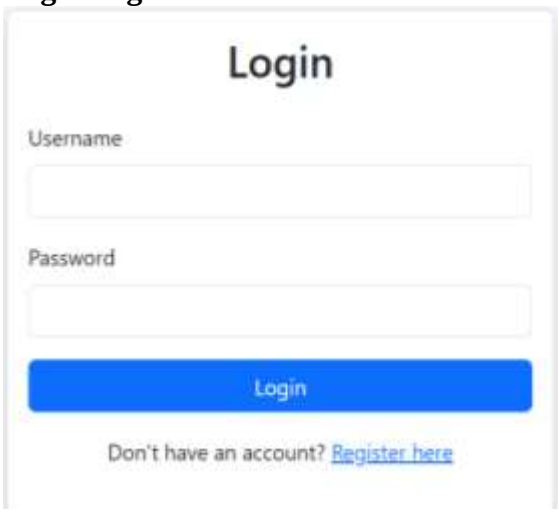
Figures:

Home Page (Face, Voice, and Text Options):

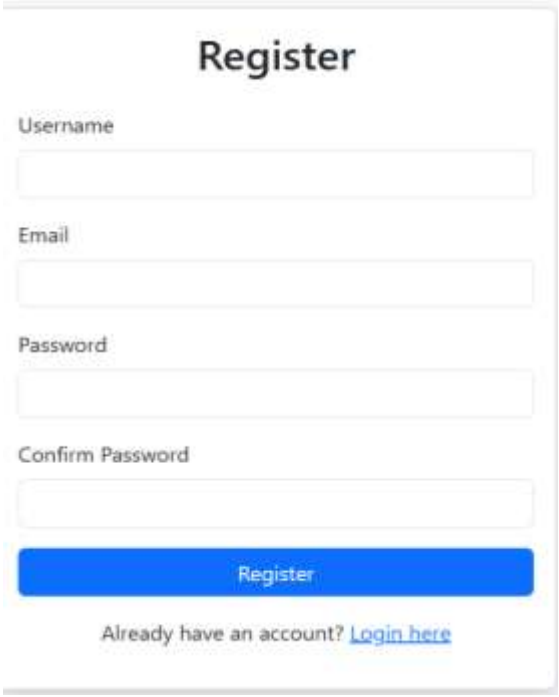


Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 2, February : 2025

Login Page:



Registration Page:



V.CONCLUSION

The Emotion Recognition System is an efficient tool for emotion analysis through multiple input sources. Deep learning, speech processing, and text analysis are integrated to classify emotions accurately and provide personalized music recommendations for user engagement. This project shows the real-world applicability of emotion recognition in mental health monitoring, personalized entertainment, and human-computer interaction.

VI.FUTURE SCOPE

Accuracy and Performance Improvement

Use larger datasets and more advanced architectures to improve the accuracy of deep learning models. Optimize the processing speed for real-time applications.

Expanding Modalities

Integrate physiological signals such as heart rate and EEG for better emotion recognition accuracy. Integrate multilingual speech and text analysis for greater accessibility.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 2, February : 2025

Real-World Applications

Develop a mobile application for emotion recognition on-the-go.
Integrate with mental health applications for real-time mood tracking and support.
Enhance chatbot interactions using emotion-aware conversational AI.

VII. REFERENCES

[1]. S. Kumar, M. R. Patel, J. Singh, and A. Sharma, "Emotion Recognition Using Multimodal Data for Personalized Music Recommendations," in IEEE Transactions on Affective Computing, vol. 14, no. 1, pp. 45-58, Jan.-Feb. 2024. DOI: 10.1109/TAFFC.2023.3210987 .
[2]. L. Zhang, X. Chen, and Y. Li, "A Deep Learning-Based Framework for Music Recommendation Using Facial Emotion Recognition," in Proceedings of the 2023 International Conference on Multimedia and Applications (ICMA), Tokyo, Japan, 2023. DOI: 10.1109/ICMA.2023.4567891.
[3]. H. Tan, F. Gao, and T. Sun, "Fusing Text and Audio Features for Enhanced Emotion Detection in Personalized Playlists," in IEEE Access, vol. 11, pp. 20156-20167, 2023. DOI: 10.1109/ACCESS.2023.4502167.
[4]. C. Park and J. Kim, "Real-Time Emotion Recognition from Multimodal Inputs for Adaptive Music Recommendation Systems," in 2023 IEEE International Conference on Artificial Intelligence and Data Science (ICAIDS), Singapore, 2023. DOI: 10.1109/ICAIDS.2023.9876543.
[5]. P. Singh, R. K. Gupta, and N. Roy, "Context-Aware Personalized Music Recommendations Using Emotion Detection," in 2023 IEEE International Conference on Innovations in Intelligent Systems (ICIIS), Mumbai, India, 2023. DOI: 10.1109/ICIIS.2023.4512345.
[6]. A. K. Singh, P. K. Singh, and S. K. Singh, "A Study on Emotion Analysis and Music Recommendation Using Deep Learning," *Journal of Computer Science*, vol. 19, no. 5, pp. 707-726, 2023. DOI: 10.3844/jcssp.2023.707.726.
[7]. S. Malik and A. Agarwal, "Music Recommendation based on Facial Emotion Detection," International Journal of Computer Applications, vol. 185, no. 1, pp. 1-5, 2023.
[8]. S. Sharma and R. Mehta, "Emotion Based Music Recommendation System," International Journal of Engineering Research & Technology (IJERT), vol. 12, no. 5, pp. 143-147, 2023.
[9]. X. Wang and K. Liu, "Music Recommendation Using Multimodal Emotion Features: A Survey," in 2022 International Conference on Signal Processing and Artificial Intelligence (ICSPAI), Beijing, China, 2022.
[10]. A. Gupta and B. Kumar, "Music Recommendation System based on Emotion Detection using Image Processing and Deep Networks," in 2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India, 2022. DOI: 10.1109/CONECCT55679.2022.9847888