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

THE REAL AUDIO EMOTION DETECTION SYSTEM

[1]Sharique Ahmad , [2]Rutik Bodke , [3]Shamoon Ansari ,[4]Bhavesh Kumar Jha , [5]Ronit Gavali

[1]Assistant Professor, [2-5] UG Student
[1] Department of Computer Engineering,
[1]Universal College of Engineering, Mumbai, India

Abstract : *This paper introduces a method for detecting emotions by using the audio data, which holds the significant importance in many real life applications such as mental health monitoring, education and learning, sentiment analysis, customer service but is has high importance in the businesses seeking to understand market insights.Detecting the emotion from the has wide range of scope. By integrating Convolutional Neural Networks (CNNs), our approach provides a comprehensive understanding of emotional response within an audio content, thereby aiding market analysis and consumer sentiment assessment. In today's era businesses increasingly rely on consumer feedback and market insights to make more informed decisions. Emotion detection from audio data offers a nuanced understanding of consumer sentiment, enabling businesses to gauge customer satisfaction, identify emerging trends and tailor their products and services accordingly. The CNN component is used to identify spatial patterns within spectrograms. The CNNs analyze the spectrogram frequency, time, amplitude to identify the spatial patterns in the audio. This helps in accurate emotion detection. Our approach helps businesses to make more informed decision making by analyzing the market. Emotion detection helps identify sentiment trends and sentiment shifts, enabling proactive reputation management strategies.By analyzing customer reactions in focus groups, product demonstrations, or user testing sessions, businesses can identify features that evoke positive emotions and areas for improvement. The result of our system on the business perspective is, it shows emotion percentage from the input data and it can be visualized through the pie-chart based on the model and the effectiveness of that product or campaign, advertisement, etc shown with the increase or decrease line graph.*

KeyWords: python Programming, Speech ,Convolutional Neural Networks, Deep Learning.

I. INTRODUCTION

In today's digital world, blending technology with human emotions is an exciting area of discovery. This technology can change how we communicate with computers. Emotions are something everyone understands, so we want to teach computers to recognize and provide the insights instantly. In the history of audio emotion detection it starts from understanding and analyzing audio signals for emotion recognition by understanding the features like pitch, intensity, then after these machine learning techniques are used to understand the emotion from the audio and now these recent years the deep learning based techniques are used to understand audio emotion. Some challenges associated with businesses before the audio emotion detection system are limited understanding of customer emotions, lack of personalized customer interactions, difficulty in identifying customer pain points, ineffective customer service responses, difficulty in competing in saturated markets to address these challenges, this paper introduces a method for detecting emotions by using the audio data, which holds the significant importance in many real life applications such as mental health monitoring, education and learning, sentiment analysis, customer service but is has high importance in the businesses seeking to understand market insights. It has more benefits in the business such as enhanced customer understanding, improved customer service, personalized marketing, product development and innovation, brand reputation, etc. Detecting emotion has a wide range of scope. By integrating Convolutional Neural Networks (CNNs), our approach provides a comprehensive understanding of emotional response within an audio content, thereby aiding market analysis and consumer sentiment assessment. Our project only takes multiple audio inputs then analyzes and after analyzing the shows the percentage of different emotions from the input audio data then visualizes it through the pie chart and based on the result it predicts whether the companies products sales can grow up or not by presenting a positive or negative line graph. Basically it shows the effectiveness of that product or campaign, advertisement, etc shown with the increase or decrease line graph.

II. LITERATURE SURVEY

Previous research has primarily focused on extracting results from unimodal systems. These systems often predict emotions based solely on facial expressions or vocal sounds. However, more recently, multimodal systems that combine multiple features have proven to be more effective and accurate in predicting emotions. These multimodal approaches incorporate features such as audio-visual expressions, EEG signals, and body gestures.To implement emotion recognition systems, researchers have employed intelligent machines and neural networks. Multimodal recognition methods, as demonstrated by Shiqing et al., outperform unimodal systems. Deep neural networks, capable of generating discriminative features that capture complex non-linear dependencies between original features, have been successfully applied to speech and language processing, as well as emotion recognition tasks.

Martin et al. highlighted the effectiveness of bidirectional Long Short-Term Memory (BLSTM) networks compared to conventional SVM approaches[1]. In speech processing, Ngiam et al. proposed and evaluated deep networks for learning audio-visual features from spoken letters. Brueckner et al.found that using a Restricted Boltzmann Machine (RBM) prior to a two-layer neural network with fine-tuning significantly improved classification accuracy in the Interspeech automatic likability classification challenge.

Stuhlsatz et al.took a different approach, using Generalized Discriminant Analysis (GerDA) based on Deep Neural Networks (DNNs) to learn acoustic features for speech emotion recognition[2]. Additionally, Yelin et al.demonstrated that three-layered Deep Belief Networks (DBNs) outperformed two-layered DBNs in audiovisual emotion recognition processes.

Previous research primarily concentrated on unimodal systems, which predict emotions based on single features (e.g., facial expressions or vocal sounds). However, more recent work has explored multimodal systems that combine multiple features for improved accuracy and stability.

Multimodal Solutions: These methods offer better performance, accuracy, reasonability, and precision. Some prioritize accuracy, while others focus on realism. Trade-offs exist, with some methods requiring more computation power but delivering higher accuracy, while others sacrifice accuracy for better performance.

Yelin Kim and Emily Mower Provos: They investigated whether a subset of an utterance could be used for emotion inference. Their proposed windowing method identifies optimal window configurations, duration, and timing for aggregating segment-level information to infer emotions at the utterance level[3]. Experimental results using IEMOCAP and MSP-IMPROV datasets demonstrated consistent patterns across speakers, specific to different emotion classes and modalities. Their method outperformed baseline approaches, even when using only 40–80% of the information within each utterance.

Consistent Patterns: The identified temporal windows consistently reveal emotion across speakers, aligning with psychological findings.

-HOLONET Method: A. Yao, D. Cai, P. Hu, S. Wang, L. Shan, and Y. Chen proposed the HOLONET method for video-based emotion recognition[4]. Key considerations in their network design include:

1. Using modified Concatenated Rectified Linear Unit (CReLU) instead of ReLU in lower convolutional layers to reduce redundant filters and enhance non-saturated non-linearity.
2. Combining residual structure and CReLU in middle layers to maintain efficiency while benefiting from increased network depth.
3. Designing the topper layers as a variant of the inception-residual structure to introduce multi-scale feature extraction.

Realistic Approach: HOLONET prioritizes adaptability in real-time scenarios over theoretical accuracy. Although its accuracy is impressive, it is specifically applicable to *video-based emotion recognition and may not work well with other data types .

CNN-LSTM and C3D Networks: Y. Fan, X. Lu, D. Li, and Y. Liu proposed a method for video-based emotion recognition in the wild[5]. They simultaneously modeled video appearances and motions using CNN-LSTM and C3D networks. The combination of these networks yielded impressive results, demonstrating the method's effectiveness .

Method and Accuracy: The proposed method achieved an impressive 85.01% accuracy. It relies on transfer learning and comprises three pillars TCA-based Subject Transfer, KPCA-based Subject Transfer, and Transductive Parameter Transfer.

-Data Preprocessing and Feature Extraction:

-Raw EEG signals were preprocessed using a bandpass filter between 1 Hz and 75 Hz.
- Differential entropy (DE features were extracted.

However, this achievement is limited to positive emotion recognition.

III. OBJECTIVE AND SCOPE

The objective of Real Audio Emotion Detection is to develop a system that can accurately identify and classify human emotions from real-time audio signals. This is a significant advancement over traditional Speech Emotion Recognition (SER) systems, which typically analyze pre-recorded audio files. By processing audio signals in real-time, Real Audio Emotion Detection systems can provide immediate feedback.Emotion plays a crucial role in human communication, influencing not only what we say but how we say it. By recognizing the emotional content of speech, Real Audio Emotion Detection systems can respond in ways that are more attuned to the user's emotional state, leading to more natural and engaging interactions. However, accurately detecting emotions in real-time audio signals is a challenging task. Unlike text, which conveys meaning primarily through words, audio signals convey meaning through a combination of words, tone of voice, pitch, volume, and timing. These acoustic features can vary widely between individuals and even between different utterances by the same individual, adding to the complexity of the task.Despite these challenges, the potential benefits of Real Audio Emotion Detection are enormous. By enabling machines to understand and respond to human emotions, Real Audio Emotion Detection can lead to more natural and engaging human-computer interactions.

The scope of Real Audio Emotion Detection encompasses a wide range of disciplines and technologies. It involves fields such as signal processing, machine learning, psychology, and linguistics, and requires a multidisciplinary approach to solve the complex challenges it presents.In terms of signal processing, Real Audio Emotion Detection involves the analysis of audio signals to extract relevant features. This requires a deep understanding of the acoustic properties of speech and the technical skills to process and analyze audio data.In terms of machine learning, Real Audio Emotion Detection involves the development of algorithms that can learn from data and make predictions. This requires a deep understanding of various machine learning techniques and the ability to apply them to the task of emotion recognition.In terms of psychology, Real Audio Emotion Detection involves the study of human emotions and how they are expressed and perceived. This requires a deep understanding of the psychological aspects of emotion and the ability to translate this understanding into a form that can be processed by a machine.

In terms of linguistics, Real Audio Emotion Detection involves the study of language and how it is used to convey emotion. This requires a deep understanding of the nuances of language and the ability to analyze and interpret linguistic data.The scope of Real Audio Emotion Detection also extends to the development of practical applications. This involves identifying potential use cases, developing prototypes, testing and refining the system, and ultimately deploying it in a real-world setting. This requires a deep understanding of the specific requirements of each application and the ability to translate the theoretical aspects of Real Audio Emotion Detection into practical solutions.

IV. PROPOSED SYSTEM

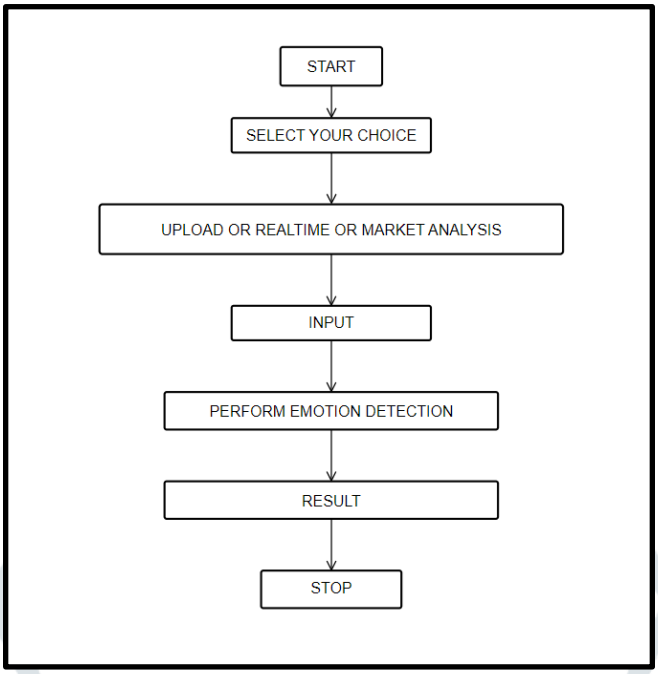

Fig.Proposed System

The proposed system offers a comprehensive platform for emotion detection from audio inputs, featuring three distinct functionalities: upload, realtime, and market analysis. The upload option allows users to submit pre-recorded audio files, prompting them to specify the conveyed emotions. Realtime functionality enables immediate emotion analysis by capturing audio input from the microphone in real-time. Market analysis caters to businesses, facilitating the assessment of customer sentiment from provided audio files. In each scenario, the system processes the audio data to detect emotions, presenting the results through structured formats or visualizations like pie charts. Whether for individual audio analysis, real-time emotion monitoring, or market sentiment assessment, this system provides a versatile solution adaptable to diverse user needs.

V. SYSTEM ARCHITECTURE

The below diagram shows the system architecture of our deep learning model project.

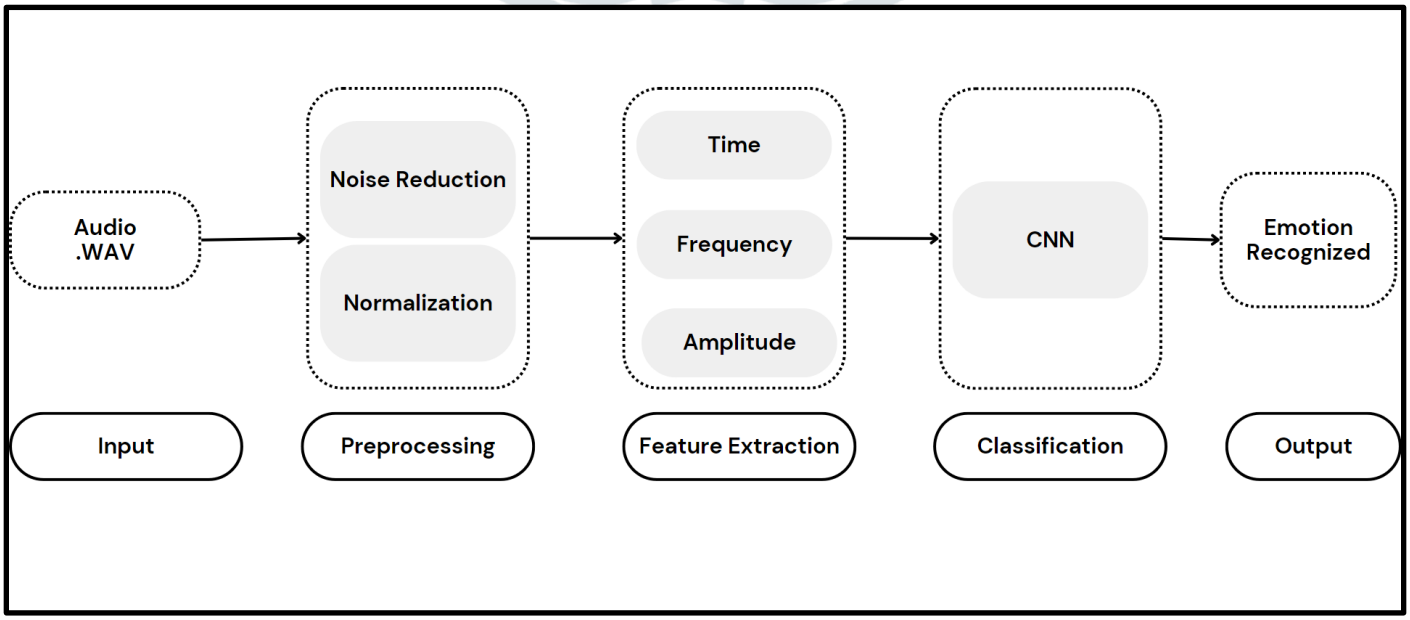

Fig.System architecture

The system takes an audio signal, which can be a .WAV file, as input. In the preprocessing stage, the system applies noise reduction techniques to remove background noise from the audio signal. It also performs normalization to ensure the volume is consistent

throughout the recording. In the feature extraction step, the system extracts features from the preprocessed audio signal. These features include time-domain features like how the audio signal changes over time, frequency-domain features like the pitch of the voice, and amplitude-domain features like how loud the audio signal is.Finally, a Convolutional Neural Network (CNN) analyzes the extracted features and classifies the emotions in the speech. The CNN outputs the recognized emotion, which could be happiness, sadness, anger, or any other emotion.

Here's a detailed explanation of the each stage:

Input:

This is the stage where the raw audio signal, in Waveform Audio Format (WAV), is fed into the system. The audio signal can be from various sources like microphone recordings or audio files stored on a computer.

Pre-Processing:

The initial step involves preprocessing the input audio signal, which is often provided in the .WAV format. During this stage, several techniques are applied to enhance audio quality:

- Noise reduction: Unwanted background noise is removed to ensure that the relevant speech features stand out.
- Normalization: It ensures that audio signals consistent volume throughout this process. This ensures quieter or louder parts don't overwhelm the feature extraction stage.

Feature Extraction:

This stage focuses on extracting relevant characteristics from the preprocessed audio.

The system extracts three types of features:

Time-domain features: These capture how the audio signal changes over time. For instance, they might track variations in volume over time.

Frequency-domain features: These focus on the frequency content of the audio signal, which is related to the pitch of the voice.

Amplitude-domain features: Here, the system extracts information about how loud the audio signal is.

Classification:

The Convolutional Neural Network (CNN) to analyze the features extracted in the previous stage. A CNN is a type of artificial intelligence particularly suited for identifying patterns. In this case, the CNN analyzes the extracted features and attempts to recognize the emotion conveyed in the speech sample. The output layer of the CNN corresponds to the recognized emotions, such as happiness, sadness, or anger.

VI. METHODOLOGY

1 Tools Utilized:

1.1 Kaggle: Kaggle is an online platform that hosts data science and machine learning competitions, providing a collaborative environment for data scientists, researchers, and enthusiasts to tackle real-world problems through innovative solutions. Founded in 2010, Kaggle offers a diverse range of datasets and challenges, spanning various domains such as healthcare, finance, and computer vision. Participants compete to develop the most effective predictive models or analytical approaches, often leveraging advanced algorithms and techniques. Beyond competitions, Kaggle serves as a hub for knowledge sharing, fostering a vibrant community through forums, tutorials, and datasets. With its vast repository of data and resources, Kaggle has become a cornerstone of the data science community, driving innovation and collaboration in the field.

1.3 Google Colab: Google Colab, short for Google Colaboratory, is a cloud-based platform offered by Google that provides free access to computing resources, including GPUs and TPUs, for writing, executing, and sharing Python code collaboratively. It integrates seamlessly with Google Drive, allowing users to create and share Jupyter notebooks with ease. With its powerful features and extensive library support, Google Colab is widely used by researchers, students, and professionals for various tasks such as data analysis, machine learning experimentation, and scientific research, making it a versatile and accessible tool for computational tasks without the need for expensive hardware or setup.

2 Dataset Utilized Description:

Selecting a suitable emotional speech dataset stands as a critical aspect of Speech Emotion Recognition (SER) design. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset is widely acknowledged as a comprehensive collection of recordings encompassing emotional speech and song, presenting a multi-modal approach to studying emotions. Emerging from 24 professional actors, this gender-balanced database features 104 distinct vocalizations portraying various emotions including happiness, sadness, anger, fear, surprise, disgust, calmness, and neutrality.Audio Processing Libraries: Libraries such as Librosa and PyDub can be used for audio processing tasks like reading audio files, extracting features, and converting between different audio formats.

3 Feature Extraction Techniques:

In real audio emotion detection systems, feature extraction plays a crucial role in transforming raw audio signals into a format that machine learning models can understand and analyze effectively. Here are some common feature extraction techniques used in such systems:

3.1 Mel-Frequency Cepstral Coefficients (MFCCs):
They represent the short-term power spectrum of a sound by extracting features based on the human auditory system's response to different frequencies. MFCCs capture both the frequency and amplitude characteristics of audio signals, making them suitable for tasks like speech recognition and emotion detection.

3.2 Spectrogram:
It is computed by applying the Fourier transform to short, overlapping segments of the audio signal.Spectrograms provide information about the distribution of energy across different frequency bands over time, making them useful for analyzing dynamic audio content, such as speech and music.

3.2 Zero-Crossing Rate (ZCR):
ZCR measures the rate at which the audio signal changes sign, indicating the presence of rapid changes in amplitude. It is often used as a feature for detecting speech and music boundaries.
In emotion detection, ZCR can capture aspects of the signal related to its temporal dynamics, such as the rate of speech or vocal fluctuations associated with emotional expression.

3.3 Energy:
Energy measures the total power in an audio signal over a given time interval. It is calculated by summing the squared values of the signal samples.Energy can be a useful feature for distinguishing between different types of sounds or determining the intensity of vocal expressions in emotion detection tasks

3.4 Pitch:
Pitch represents the perceived frequency of a sound and is closely related to the fundamental frequency of a periodic signal.
Extracting pitch-related features can help capture characteristics such as prosody and intonation in speech, which are important cues for inferring emotional states.

3.5 Formant Frequencies:
Formants are resonant frequencies in the vocal tract that contribute to the distinctive timbre of speech sounds. Extracting formant frequencies can provide information about the shape and dynamics of the vocal tract during speech production.
Formant-based features are particularly relevant for emotion detection in speech, as changes in vocal tract configuration are associated with emotional expressiveness.

3.6 Temporal and Frequency Domain Statistical Features:
These include statistical measures such as mean, standard deviation, skewness, and kurtosis computed over time or frequency bins.Statistical features capture higher-order characteristics of the audio signal distribution and can provide insights into its variability and shape, which are relevant for discriminating between different emotional states.

4 Deep Learning Techniques Used :

4.1 Convolutional Neural Networks:
Convolutional Neural Networks (CNNs) are widely used in Real Audio Emotion Detection due to their ability to process and learn from both spectral and temporal features of audio signals.In Real Audio Emotion Detection, the audio signal is first converted into a spectrogram, which is a 2D representation of the signal where one axis represents time, and the other represents frequency. The output of each layer is then passed through a non-linear activation function, which introduces non-linearity into the model, allowing it to learn more complex patterns. The output is also typically downsampled using a pooling layer, which reduces the dimensionality of the output and helps to make the model more robust to variations in the input.After several such layers, the output is flattened into a 1D vector and passed through one or more fully connected layers, which perform the final classification of the emotion.The CNN is trained using a large dataset of labeled audio signals, where the label indicates the emotion expressed in the signal. During training, the weights of the filters in the CNN are adjusted to minimize the difference between the predicted and actual labels.In this way, CNNs can be used to detect emotions in real audio signals, providing a powerful tool for Real Audio Emotion Detection.

VII. RESULT AND DISCUSSION

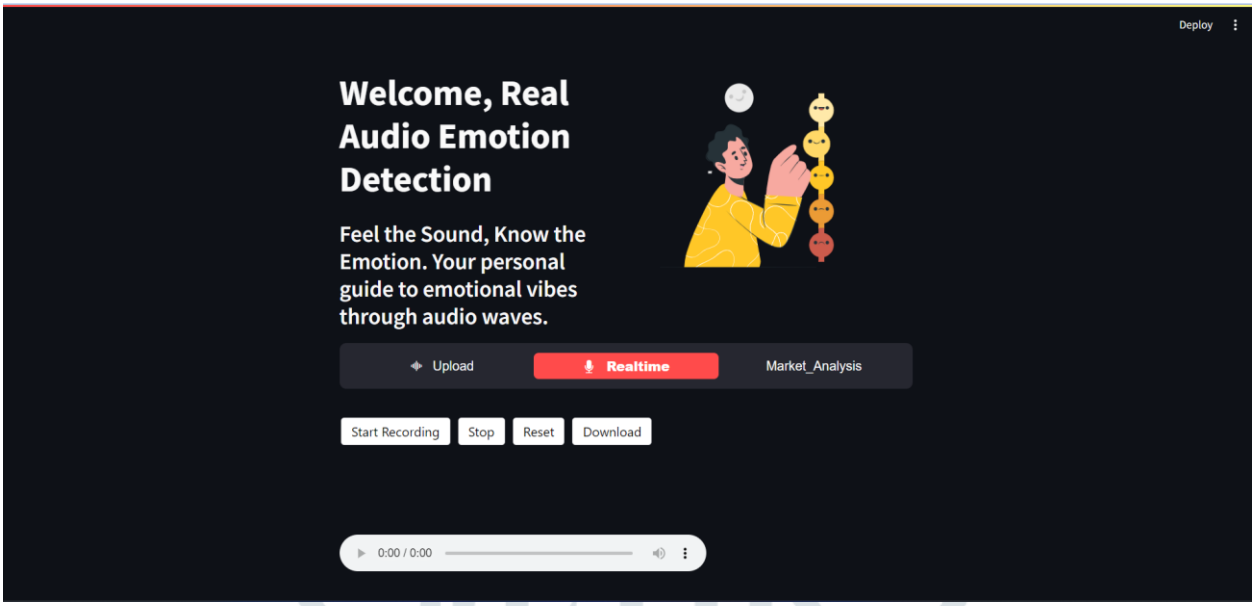

Fig.User Interface

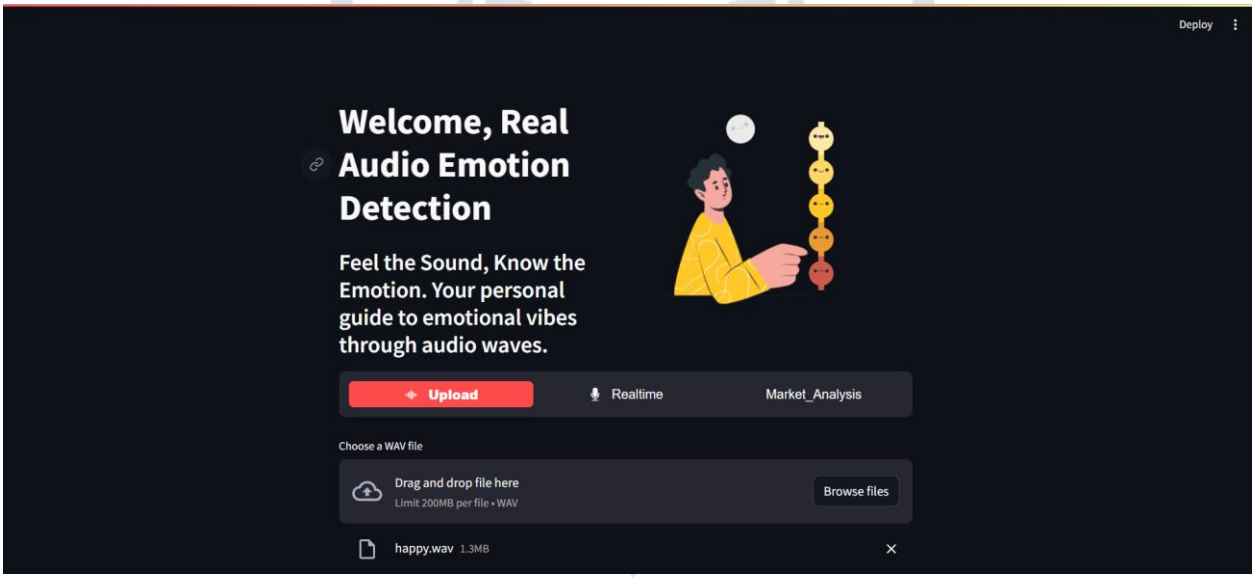

Fig.Using Upload Section

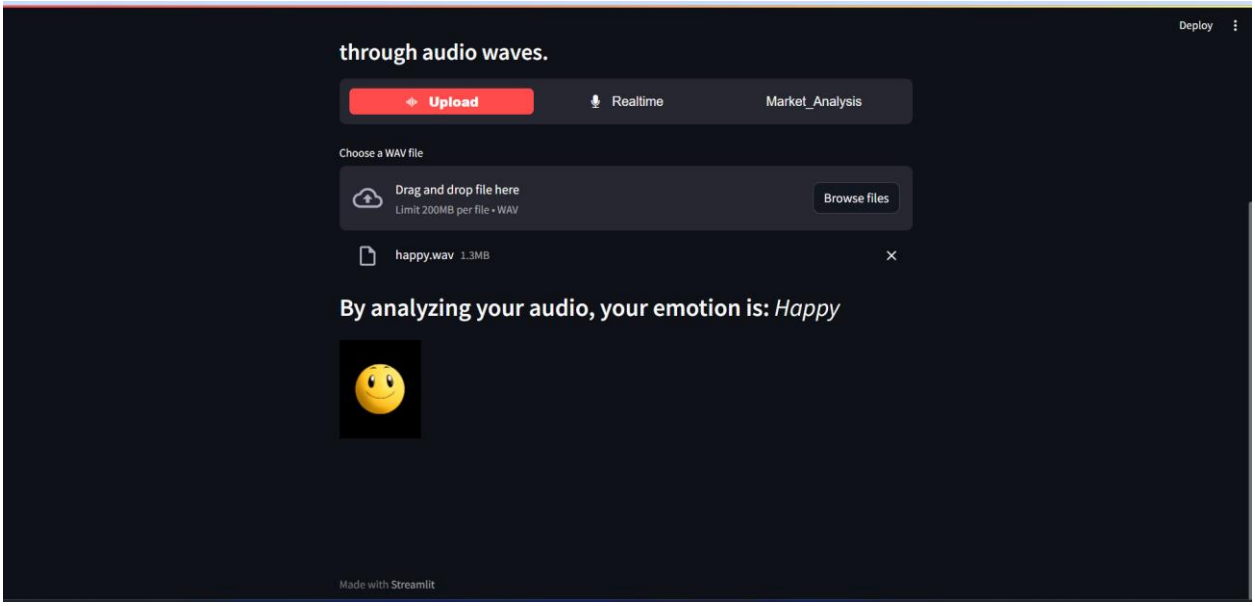

Fig.Output of Upload Section

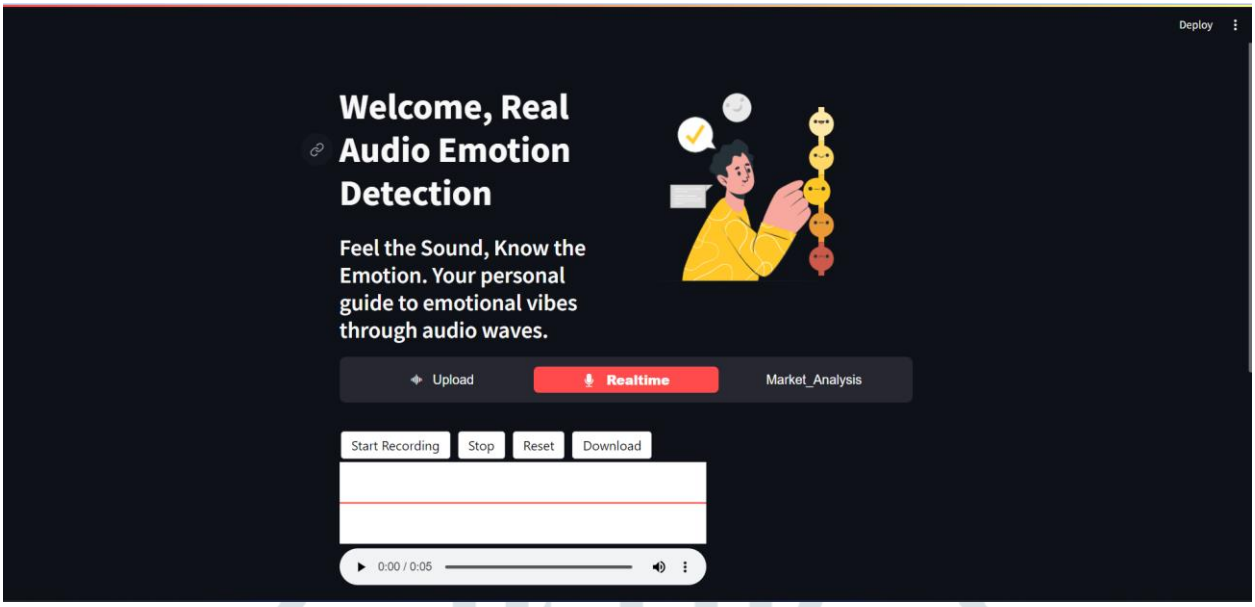

Fig.Using Real Time Section

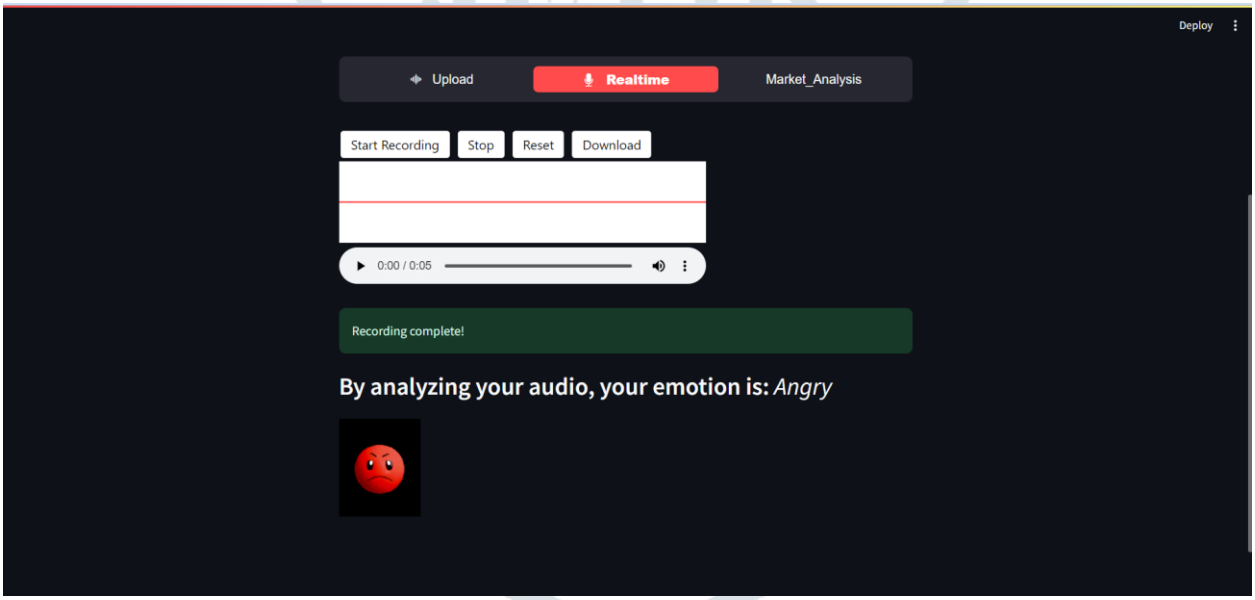

Fig.Output Of Real Time Section

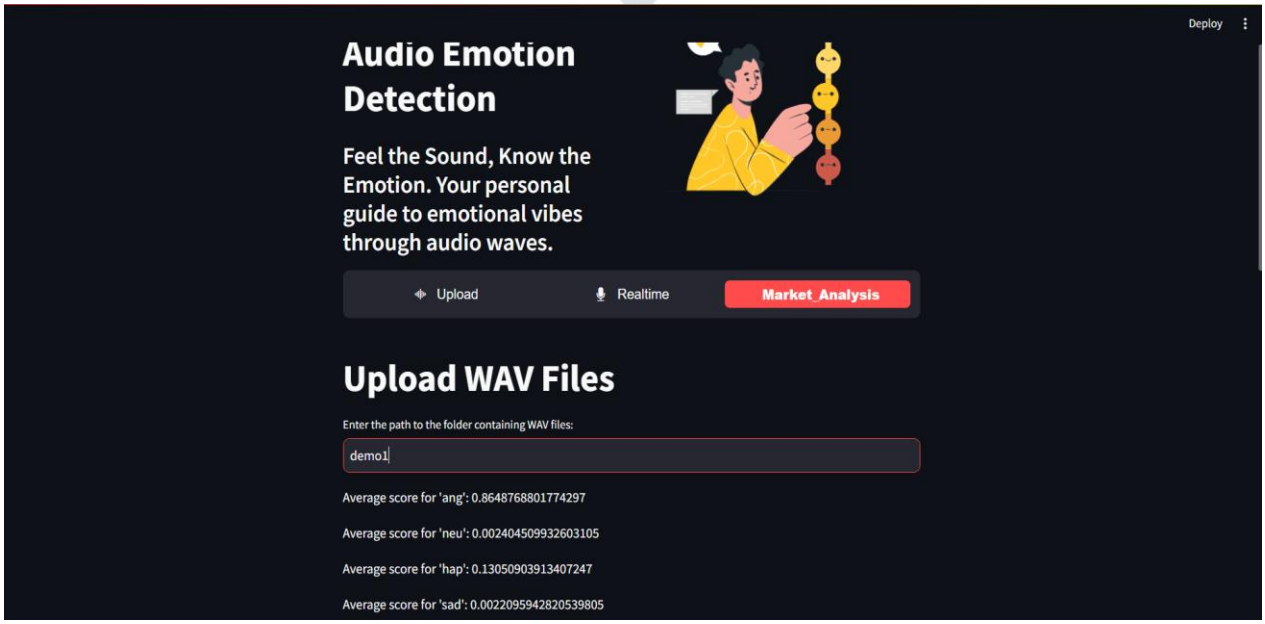

Fig.Providing Input Files Folder in MarketAnalysis

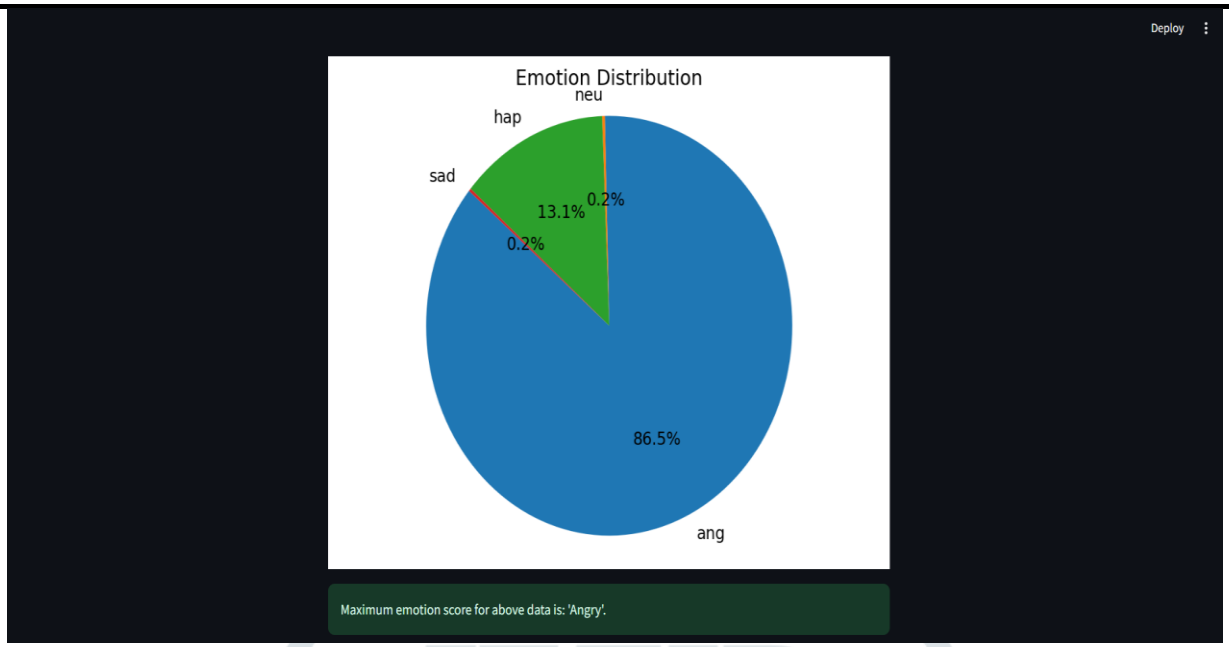

Fig.Output of MarketAnalysis with visualization

VIII. CONCLUSION

This study makes a significant contribution to the field of emotion recognition by introducing the Real Audio Emotion Detection System. The system leverages Convolutional Neural Networks (CNNs) and deep learning techniques to provide valuable insights into consumer sentiment and behavior. By analyzing and visualizing emotion percentages through pie charts and line graphs, the system offers actionable insights for proactive reputation management and informed decision-making.It takes an input audio signal, which can be in .WAV format. During preprocessing, noise reduction and normalization techniques are applied to enhance audio quality. In the feature extraction step, relevant features such as time, frequency, and amplitude are extracted from the preprocessed audio. Finally, CNN analyzes these features and predicts the conveyed emotion in the speech. The output corresponds to recognized emotions, such as happiness, sadness, or anger.

In the market analysis section, business people can see if the product will be successful by using real-time audio detection. A pie chart shows the number of happy and unhappy customers. The line chart will also show if the product will be good or bad in the market.With this information, businesses can improve their strategies to make the product better and more successful in the market and this will also help the business to increase their profit.

IX. REFERENCES

[1] Kaur and Singh, “speech emotion recognition”, Institute of Electrical and Electronics Engineers (IEEE), 2023, Vol. 8, No. 4, pp. 90-96.

[2] SB Shah, “Emotion recognition in speech”, Institute of Electrical and Electronics Engineers (IEEE), 2023, Vol. 9, No. 3, pp. 29-36.

[3] Kamaldeep Kaur, “Human computer interaction”, Institute of Electrical and Electronics Engineers (IEEE), 2023, Vol. 9, No. 8, pp. 49-58.

[4] Ram Sahu, “Speech Emotion Recognition Systems”, Institute of Electrical and Electronics Engineers (IEEE), 2023, Vol. 7, No. 4, pp. 90-98.

[5] Krishi Sanskriti, “Speech Emotion Recognition Systems”, Institute of Electrical and Electronics Engineers (IEEE), 2022, Vol. 2, No. 3, pp. 19-26.

[6] S Tripathi, “Deep Learning based Emotion Recognition System”, Institute of Electrical and Electronics Engineers (IEEE), 2022, Vol. 6, No. 5, pp. 60-76.

[7] Samesh Madanian, “Speech emotion recognition using machine learning”, Institute of Electrical and Electronics Engineers (IEEE), 2022, Vol. 11, No. 6, pp. 80-96.

[8] Anuj Yadav, “Speech Emotion Recognition and Vision Systems Using a Recurrent Neural Network (RNN)”, Institute of Electrical and Electronics Engineers (IEEE), 2022, Vol. 12, No. 7, pp. 100-126.

[9] Kranthi Sai Reddy Vanukuru, “Emotion Detection Through Audio Using Machine Learning”, Institute of Electrical and Electronics Engineers (IEEE), 2021, Vol. 15, No. 8, pp. 290-296.

[10]Dr.Smith's,“Real Audio Emotion Detection ”,Institute of Electrical and Electronics Engineers(IEEE), 2021,Vol. 9, No. 9, pp. 470-286.