National College of Ireland

Voice to Value: Multimodal Emotion Analysis for Smarter Customer Interactions

Research Practicum

Bharathi Ramapatnam
Student ID: x23425237

School of Computing
National College of Ireland

Supervisor: Arjun Chikkankod

National College of Ireland
Project Submission Sheet
School of Computing

Forename Surname:	Bharathi Ramapatnam
studentID:	X234225237
Programme Name:	Masters In Data Analytics
Year	24-25
MSc Research Project	Research Practicum
Supervisor:	Arjun Chikkankod
duedate:	11-08-2025
Title	Voice to Value: Multimodal Emotion Analysis for Smarter Customer Interactions
Word Count:	5210
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Bharathi Ramapatnam
Date:	14th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Voice to Value: Multimodal Emotion Analysis for Smarter Customer Interactions

Bharathi Ramapatnam
x23425237

Abstract

Sentiment analysis has been a major focus in natural language processing, traditionally centered on text. However, in voice-driven interactions across healthcare, education, and customer service, text alone often misses emotional cues present in speech. This research introduces a multimodal sentiment analysis framework that independently analyzes speech emotion and text sentiment to capture a fuller emotional profile. Deep learning models—including 1D and 2D Convolutional Neural Networks (CNNs), Random Forest, and Recurrent Neural Networks (RNNs)—are applied to the RAVDESS and EMOGATOR datasets. Audio features are extracted using Mel-frequency cepstral coefficients (MFCCs), while textual sentiment is analyzed using transformer-based models. The 2D CNN outperforms baseline models, achieving 71% accuracy on RAVDESS and 63% on EMOGATOR, compared to 61% from Random Forest. AssemblyAI's speech-to-text API enables accurate transcription, with the full system integrated into a web application for real-time emotion tracking. A total of 13 emotion classes are classified. Results show that parallel analysis of audio and text provides complementary insights into user sentiment. This approach offers practical benefits for customer relationship management (CRM) systems and healthcare platforms, enabling more responsive and personalized services based on emotional understanding.

1 Introduction

There are several emotions that audio can capture but are often missed when relying purely on text. For example, emotions such as anger, sadness, or sarcasm are heavily dependent on tone of voice, which is typically lost during transcription. Based on Dennison's framework Dennison (2024a) for audio sentiment analysis by mapping Plutchik's 32 emotions to vocal features like pitch, tone, and tempo. His emphasis on emotional framing and physiological reactions helps guide which emotions to target based on behavioral outcomes. Pitch and speech rate—such as a rising or falling tone—can indicate whether a person is expressing doubt, excitement, or asking a question. These minor vocal cues reveal behavioural patterns that are difficult to detect in text alone. Similarly, changes in volume or emphasis on specific words carry emotional weight that plain text fails to convey. Several sectors could benefit from this type of analysis. A recent implementation in the healthcare sector and emotion recognition on call voice data based on Turkish voice calls in call center setup Alhussein et al. (2025),Yurtay et al. (2024). Extracting both agents and clients voice data for sentiment analysisPłaza et al. (2022) are

the evidence of needs for customized sentiment analysis(SA) in call center setup. Phone-based healthcare services have become increasingly vital, especially in the post-pandemic era, where remote consultations, mental health support, and chronic disease management are now standard practices. These interactions often rely exclusively on verbal communication, making it critical to understand the emotional tone and sentiment behind a patient's speech. Integrating audio sentiment analysis into telehealth systems can help providers detect emotional distress, tailor their responses, and ultimately improve patient outcomes. Video representation showing facial expressions would provide a more complete picture of a person's emotionsYi et al. (2025). However, obtaining video footage of clients is often not feasible due to data privacy concerns and ethical constraints, file size and storage limitations. In sectors such as sales, customer service, and product support, most interactions occur over the phone. While transcribing these conversations into text provides valuable information, adding audio sentiment analysis uncovers emotional cues that text alone cannot offer. This deeper emotional insight allows businesses to better understand customer needs, deliver more personalized experiences, and make data-driven decisions that enhance customer satisfaction. Considering the outcomes of the literature review across four key areas—text-based sentiment analysis, audio-based emotion recognition, bi-modal approaches, and pre-trained OpenAI models—it is evident that each has its own strengths and limitations. Upon analyzing these studies, a clear gap emerges: the need for a tailored solution that closely mimics real-time customer service or call center scenarios. This highlights the importance of developing a domain-specific, easily scalable model that can operate effectively under practical constraints while maintaining high accuracy. To address this gap, this paper proposes a lightweight, low-cost hybrid sentiment analysis model that fuses audio and text inputs for more accurate emotion recognition in call center and healthcare contexts. The solution is designed for easy implementation and real-time deployment, leveraging AssemblyAI's speech-to-text API to ensure high transcription accuracy with minimal latency. Two emotional speech datasets: RAVDESS and EMOGATOR. The approach involves exploring various audio feature extraction techniques, such as Mel-frequency cepstral coefficients (MFCC), Mel spectrograms, and chroma features. To mitigate class imbalance, data augmentation techniques like pitch alteration are applied. High-quality resampling methods, such as kaiser_best, are used to ensure consistent audio quality. Finally, a Convolutional Neural Network (CNN) is employed for multi-class classification, treating the extracted features as image-like input for emotion recognition. building a powerful hybrid model that fuses audio analysis with text-based sentiment detection. Initially, spoken content is transformed into spectrograms, which are processed by a CNN to extract emotional cues from voice patterns. Activation functions such as ReLU (Rectified Linear Unit) is chosen over Tanh, Lower learning rate of 0.00005 is adopted to maximize model learning capabilities, optimized value of RAMSprop learning rate parameter is adopted for normalization. Simultaneously, the speech is transcribed into text using Automatic Speech Recognition (ASR) and analyzed using advanced pre-trained language models such as RoBERTa and DistilRoBERTa to assess sentiment. Since the RAVDESS dataset suffers from class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is applied to synthetically augment the minority class. Overall, 13 sentiments are successfully classified using a 2D-CNN, achieving 72% accuracy on RAVDESS and 65% on EMOGATOR. Base models such as 1D-CNN and Random Forest are also implemented to compare accuracy against traditional approaches. Embedding the well-trained deep learning model into AssemblyAI, which specializes in providing state-of-the-art speech-to-text APIs, results

in 95% transcription accuracy with minimal latency. This makes it an ideal candidate for integration with systems such as Customer Relationship Management (CRM) platforms and call center software. The fusion of acoustic and linguistic features creates a robust representation of emotional and sentiment content, significantly enhancing prediction accuracy. The result is a well-rounded multi-modal solution suitable for real-world deployment. multi- model suitable for real-world applications like mental health screening, customer feedback analysis, or conversational AI. An end-to-end flow—from input, feature extraction (mel, chroma, MFCC), to emotional and textual label prediction—is tested using unseen test data. MELD (Multimodal EmotionLines Dataset) and YouTube audio are used to validate real-time performance in customer care scenarios. This approach is further supported by Khan et al. (2021), who utilized YouTube sentiment data for naturalistic opinionated audio analysis.

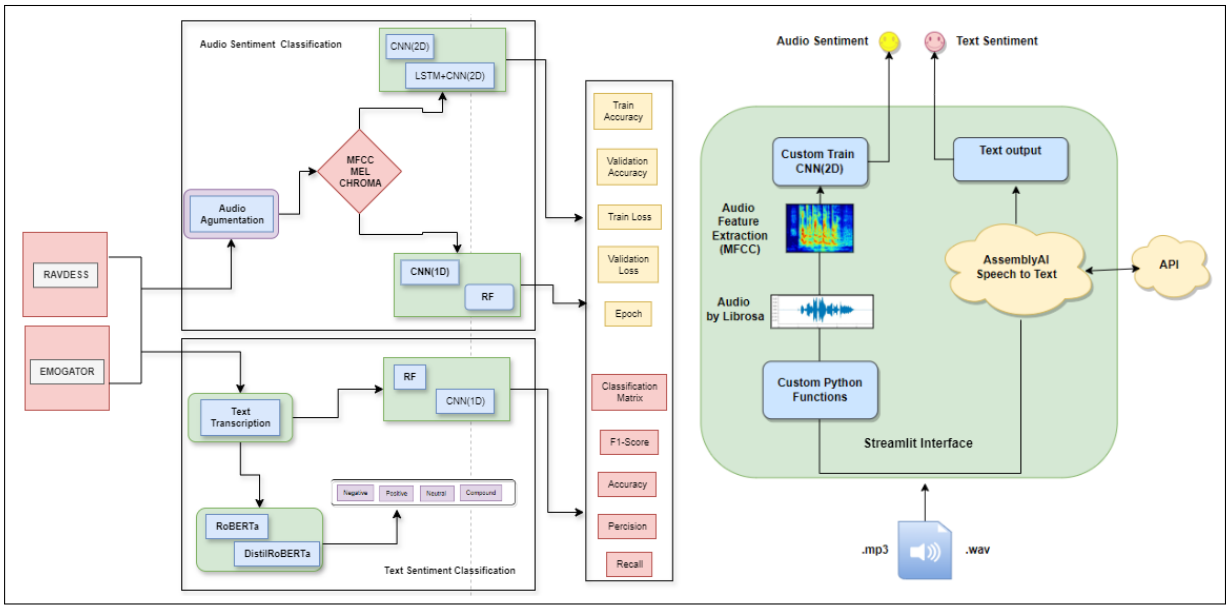

Figure 1: Overall System Architecture, Web Interface with Assembly AI And Audio Analysis

2 Related Work

2.1 State-of-the-art techniques in text sentiment analysis

The latest advancements in text-based sentiment analysis are largely driven by transformer-based models such as DeBERTa, RoBERTa, BART, and GPT-4o. These models are pretrained on massive corpora and have demonstrated superior performance across a wide range of NLP tasks. In this study, sentiment analysis is performed using DistilRoBERTa, a compressed version of RoBERTa. DistilRoBERTa is recognized for its lightweight architecture and high-speed processing capabilities, making it an efficient alternative to larger models such as GPT-4o, particularly in real-time or resource-constrained environments and budget limited environments. LangChain and fine tuned RoBERTa integration by Pathak and Rathor (2025)demonstrates how it can be integrated with LLM(Large Language Model) frameworks. Understanding customer sentiments through

textual data has been a significant area of focus in deep learning research for several years. This field has gained considerable attention due to the exponential growth of online reviews and user-generated content across e-commerce platforms and social media. A substantial body of work has been conducted to extract, interpret, and utilize customer opinions to inform business strategies and enhance user experience. Recent studies by Krishna et al. (2025),Ali et al. (2025),Sutedja and Hendry (2025) specifically explore the use of advanced text analysis techniques to evaluate customer reviews. These studies leverage state-of-the-art natural language processing (NLP) models to identify underlying sentiments and detect fraudulent behavior. Among them, one notable approach utilizes a high-performance fake review detection model based on a pretrained DeBERTa (Decoding-enhanced BERT with disentangled attention) architecture. This model demonstrates exceptional effectiveness, achieving an impressive 98.5% accuracy in distinguishing between authentic and fabricated reviews. The use of DeBERTa is particularly significant, as it builds upon the transformer-based architecture of BERT but incorporates enhancements such as disentangled attention mechanisms and improved positional encoding. These features enable more deeper understanding of language context, making the model especially well-suited for detecting re-occurring patterns often found in deceptive reviews. By accurately identifying fake reviews, such models contribute not only to improved customer trust and transparency but also to the integrity of online review systems—a critical concern for digital platforms.

2.2 The evolution of Audio based sentiment analysis

The most recent publication by Zhongliang Wei et al. (2025) was released in 2025. This study employed an architecture similar to the proposed project, utilizing the RAVDESS dataset and a CNN-based classification model. By extracting audio features using Mel spectrograms, the study achieved an overall accuracy of 80%. Although the paper claims to have reached up to 90% accuracy in identifying emotions such as 'calm', 'happy', and 'angry', it reported lower performance—between 60% and 72%—for emotions like 'sad' and 'disgust'. In contrast, the experiment conducted as part of this project successfully achieved classification accuracies of 80% and 84% for 'sad' and 'disgust', respectively. Designed for real-time sentiment detection in customer service environments, such as call centers Song and Zhou (2024), Waleed and Shaker (2025) published in June 2025, presents another successful implementation of audio classification using the RAVDESS dataset, utilizing all 7,356 files—including 1,440 audio-only files, 2,880 audio-video files, and 1,440 video-only files—in combination with the MELD dataset, which contains 13,000 multi-party conversations from a television show labeled with emotions such as neutral, sadness, surprise, anger, content, fear, and happiness. This represents extensive coverage and a wide variety of audio inputs, mimicking real-world scenarios where customer service agents handle a large volume of calls daily. However, a limitation of this approach is the significant computational demand required for audio feature extraction and subsequent data augmentation. The paper reports an impressive 91.9% accuracy on RAVDESS using a lightweight 1D Convolutional Neural Network (1D-CNN). It is important to note that these results were achieved in a controlled environment, leveraging high-speed GPUs and substantial RAM, which may not reflect performance on resource-constrained devices.

2.3 Challenges in Real time Bi-modal Approach

Considering multimodal approach from Cai et al. (2025),Yurtay et al. (2024) where audio and text features are combined to determine final emotion classification, HuBERT is employed to extract audio tokens from speech data. HuBERT (Hidden-Unit BERT), a self-supervised speech representation model developed by Facebook AI, is pre-trained on unlabeled audio and has proven effective in emotion recognition tasks. For text feature extraction, the RoBERTa pre-trained language model is used to generate text tokens. While HuBERT offers strong performance due to its ability to learn rich representations from raw audio, it is computationally intensive and not optimized for real-time applications. The objective of this research is to provide a production-friendly solution, with the aim of implementing the final system on a web platform or within a CRM environment to enable real-time emotion analysis. This requires both speed and accuracy to be optimized to the highest degree.

2.4 Production Ready Artificial Intelligence(AI) and Emotion Detection

Artificial intelligence is widely used across various sectors, as highlighted by Alhussein et al. (2025),Luitel et al. (2024),Parmanand and Anjali (2025). Generative AI has shown promising applications in the field of speech recognition, particularly in classifying real and fake voices using relatively simple model architectures. Features are typically extracted using spectrograms or Mel-Frequency Cepstral Coefficients (MFCCs), with data augmentation achieved through Generative Adversarial Networks (GANs).These studies highlight the successful implications of AI, though they also underscore certain caveats, such as increased computational cost, the need for large datasets to effectively train models, high sensitivity to audio quality, and class imbalance issues. Utilising the self-supervised transformer-based speech model developed by Facebook and used in the study by Fernandez and Awinat (2024), the authors experimented with the CREMA-D dataset for audio and the RAVDESS dataset for video. The model achieved an accuracy of 74% on the audio data, with a training time of 1 hour and 5 minutes—slightly high due to the integration of both audio and video features. Recent advancements use models like RoBERTa and DeBERTa for sentiment analysis, with DistilRoBERTa offering lightweight performance. Audio models using RAVDESS and CNNs achieve up to 91.9% accuracy, though some emotions remain harder to detect. Bi-modal approaches with HuBERT and RoBERTa enhance results but demand high computational resources. GDPR concerns arise when using platforms like OpenAI, requiring secure handling of audio data. This study addresses a key gap by using 13 emotion classes from RAVDESS and EMOGATOR, applying augmentation to balance classes, extracting MFCCs, and using 2D CNNs to improve emotion recognition in realistic scenarios.

3 Research Methodology

This research follows a structured seven-phase methodology to develop a multimodal sentiment analysis system that integrates audio and text inputs for emotion recognition. It begins with a comprehensive literature review to identify gaps in existing sentiment analysis approaches, particularly in real-time applications within customer service and

healthcare. The problem is defined as the lack of scalable, domain-specific models capable of capturing emotional nuances from speech and text simultaneously. The system design incorporates deep learning models for audio—such as 1D and 2D Convolutional Neural Networks (CNNs), BiLSTM, and Random Forest—and transformer-based models like RoBERTa and DistilRoBERTa for text sentiment analysis. Audio data is sourced from RAVDESS and EMOGATOR datasets, while MELD and YouTube samples are used for real-world validation. Preprocessing includes feature extraction using Mel-frequency cepstral coefficients (MFCCs), Chroma, and Zero Crossing Rate (ZCR), along with data augmentation techniques like pitch shifting and time stretching to address class imbalance. Speech is transcribed using AssemblyAI's Automatic Speech Recognition API, enabling real-time conversion of audio to text. The transcribed text is then analyzed using pre-trained transformer models to classify sentiment into Positive, Neutral, or Negative categories. To improve model generalization and mitigate overfitting, techniques such as SMOTE, dropout regularization, and batch size optimization are applied. The system's value lies in its ability to deliver emotionally intelligent insights in real time. By fusing acoustic and linguistic features, it enhances sentiment prediction accuracy and enables more empathetic, personalized interactions. This has practical implications for CRM platforms, telehealth services, and conversational AI, where understanding user emotions can improve engagement, trust, and decision-making. The final system, deployed via a Streamlit interface, offers a scalable, low-latency solution ready for integration into real-world applications.

4 Design Specification

4.1 Data collection and preprocessing

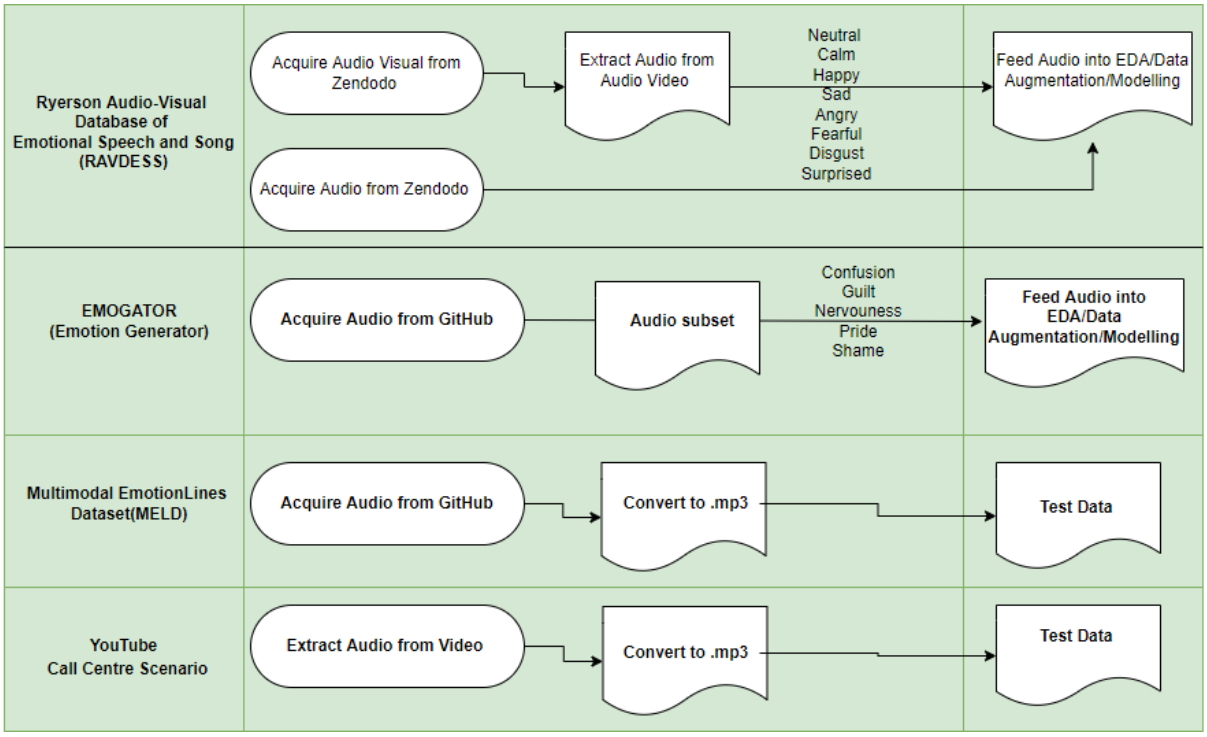

Figure 2: Audio-Based Data Acquisition Workflow from Diverse Datasets

This study utilizes three publicly available emotion-labeled datasets: RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song), EMOGATOR, and MELD (Multimodal EmotionLines Dataset). From the RAVDESS dataset, originally hosted on Zenodo and comprising 7,356 recordings, a total of 4,320 samples are selected for analysis—specifically, 1,440 audio-only speech files in .wav format and 2,880 audio-visual speech files in .mp4 format. RAVDESS is a gender-balanced dataset featuring 24 professional actors (12 male and 12 female), each expressing eight distinct emotions: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. Each actor delivers multiple renditions of two scripted sentences—"Kids are talking by the door" and "Dogs are sitting by the door"—at two emotional intensities: normal and strong. This controlled structure ensures consistent variability in both content and emotional expression, making it ideal for training and evaluating emotion recognition systems.

The EMOGATOR dataset, available under the Apache 2.0 open-source license and hosted on GitHub, contains 32,130 vocal burst samples spanning 30 emotion categories. With recordings from 357 speakers, EMOGATOR offers high diversity and is particularly suited for simulating real-world scenarios such as customer service or call center interactions. To optimize training time and manage storage constraints, this study uses a targeted subset of EMOGATOR, focusing on emotion categories not present in RAVDESS—specifically confusion, guilt, nervousness, pride, and shame. This selection broadens emotional coverage while maintaining computational efficiency.

Additionally, the MELD dataset is employed for testing purposes. MELD is an audio-visual dataset containing 13,000 audio files labeled with seven emotions: anger, disgust, sadness, joy, neutral, surprise, and fear. To further evaluate end-to-end system performance, several audio samples were acquired from YouTube, specifically from customer service and call center scenarios. These samples were used to test audio and text sentiment analysis within a Streamlit-based web application.

4.2 Model selection and justification

The goal of this study is to evaluate and compare the performance of various machine learning and deep learning models for audio and text-based emotion classification. A combined approach of traditional and deep learning models are selected based on there abilities in classifying the emotions accurately. Random Forest (RF): Used as a baseline model due to its simplicity, interpretability, and robustness on small to medium-sized datasets. It serves as a performance benchmark for comparison with deep learning approaches. 1D Convolutional Neural Network (1D-CNN) are applied to sequential features like MFCCs to capture temporal patterns in the audio signal. Suitable for lower-dimensional inputs with relatively fast training times. 2D Convolutional Neural Network (2D-CNN) is utilized to process spectrogram-like inputs treating them as images. For text based SA transformer-based models such as RoBERTa and DistilRoBERTa are used on the text extracted from audio files. This model is known for its efficiency and ease to integrate with real-time SA. Like RoBERTa, DistilRoBERTa is built on the transformer architecture, which allows it to understand contextual relationships within sentences—crucial for accurate sentiment detection, especially in short transcribed speech segments

4.3 Speech-to-Text Using AssemblyAI- Automatic Speech Recognition

To provide real-time speech-to-text transcription, AssemblyAI was chosen for its high accuracy, ease of use, and low latency. Designed to support .wav and .mp3 file formats, AssemblyAI offers a built-in solution for speech analysis (SA) through a customizable interface. By consuming the power of AI via an API (Application Programming Interface) pretrained on a large and diverse set of audio datasets, it eliminates the burden of training models from scratch. While the API itself is not customizable, it is integrated into a Streamlit-based web interface that allows users to upload audio files, preview waveforms and spectrograms, extract MFCC features, and visualize predicted emotions alongside AssemblyAI's transcription and sentiment analysis insights.

4.4 Training, Evaluation, and Data Augmentation Procedures

In machine learning tasks like audio classification, the goal is to build a model that generalizes well to unseen data. Before applying classification algorithms such as Random Forest Classifier or Convolutional Neural Networks (CNNs), the vectorized dataset is typically split into training and testing sets — commonly at ratios of 70/30 or 80/20 — to facilitate effective evaluation. In experiments using the RAVDESS dataset, a significant class imbalance was observed, notably with only 288 files corresponding to the "Calm" emotion. This imbalance can degrade model performance, especially in domains like emotion recognition by skewing predictions toward dominant classes. To address this, as prior studies such as Hemmatian et al. (2025)**?** implemented hybrid models combining SMOTE-based oversampling, feature normalization, and CNNs. These approaches proved effective in improving classification outcomes, with Random Forest accuracy increasing by 2% — highlighting the impact of balancing techniques on model reliability. To increase diversity in audio feature recognition and enhance the model's generalization capability, data augmentation techniques were employed. This approach is well supported in the research domain, notably by such as Galić and Đorđe Grozdić (2023). Overfitting during the implementation of a 2D Convolutional Neural Network (CNN) for audio classification occurred as the model because too sensitive to the training data, memorizing specific patterns, noise, and inconsistencies rather than learning features. This led to high training accuracy on the training data but poor performance on test data samples. Leveraging audio data augmentation techniques such as noise pitch shifting, and time stretching were implemented. These techniques significantly improved the CNN's performance, increasing accuracy from 65% to 72%. To assess the performance of the text and audio classification model, several metrics were employed: Accuracy measures the overall percentage of correctly predicted samples across all classes. Precision to indicates how many of the predicted positive samples were actually correct particularly in measuring false positives. Recall reflects how many of the actual positive samples were correctly identified which is to measure false negatives – when a model fails to detect correct labels. Real time impact of false negative audio class is when a user emotional expression is falsely classified as "Neutral" undermining user empathy, trust and actual state of human emotion. F1 Score: was computed to evaluate overall classification performance across all emotion categories, treating each class equally regardless of support. This approach is well-suited to multi-class problems where class imbalance may exist. Confusion Matrix: A confusion matrix was generated to visualize the distribution of predictions versus

actual labels. This matrix provides insight into model accuracy per class and highlights areas of misclassification. The plot was rendered using a blue color map with rotated x-axis labels for readability. To evaluate the dynamics of the 2D CNN model, training and validation curves were plotted across multiple epochs. The accuracy plot highlights how well the model is learning to classify the audio data, showing separate trends for both training and validation sets. The loss plot provides insight into how the model's prediction error evolved during training.

5 Implementation

5.1 Feature Extraction – Exploratory Data Analysis

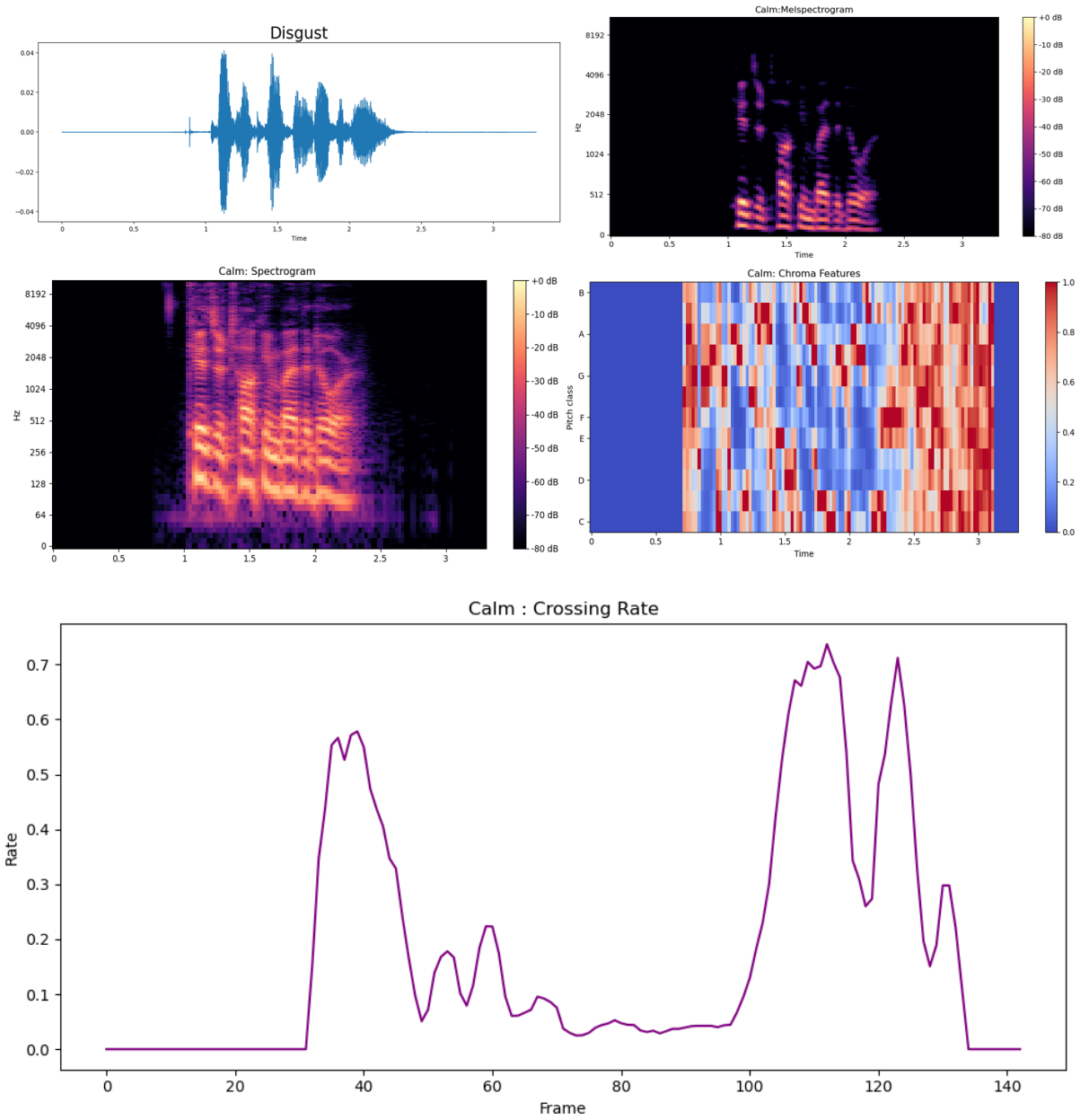

Figure 3: Audio EDA- Audio Waveform, Chroma Features, Mel Spectrogram, MFCCs, and Zero Crossing Rate(ZCR)

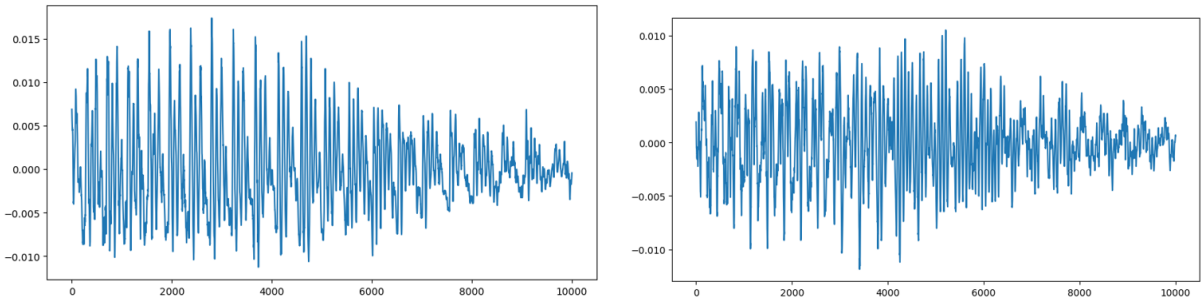

Figure 4: Pitch Shift in Audio Augmentation

Initial analysis of the RAVDESS and EMOGATOR datasets involved waveform inspection, feature extraction (MFCC, Chroma, ZCR), and signal visualization to uncover emotional distinctions. This is shown in Fig 3. Comparisons between tones like 'Disgust' and 'Calm' revealed spectral and temporal contrasts that informed architecture and feature selection. MFCCs provide compact timbral representations, capturing emotion-related pitch and vocal texture. Zero Crossing Rate(ZCR) measures how frequently the waveform crosses the zero amplitude line—higher rates can indicate excitement or aggression, while lower ones suggest calm speech. Chroma features track pitch class distributions over time, useful for identifying expressive modulations and typical tonal patterns in emotional speech.To diversify training data, time stretching was applied as shown in Fig. 4, speeding and slowing it by 20%. These increases simulate different speech rates without altering core emotional cues. Volume scaling—making audio 50% louder or quieter—adds variability in vocal intensity, helping models generalize across speakers and recording conditions. Together, these techniques enhance emotion recognition accuracy by enriching input diversity while preserving essential emotional characteristics.

5.2 Feature Extraction and Model development

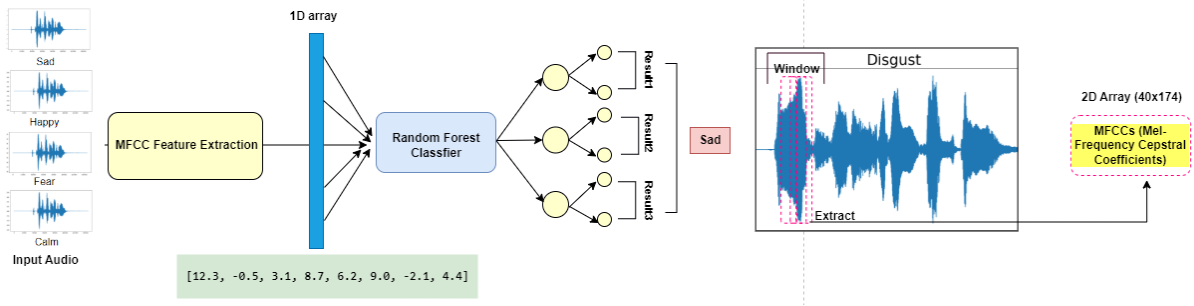

Figure 5: Random Forest and MFCC Feature Extraction

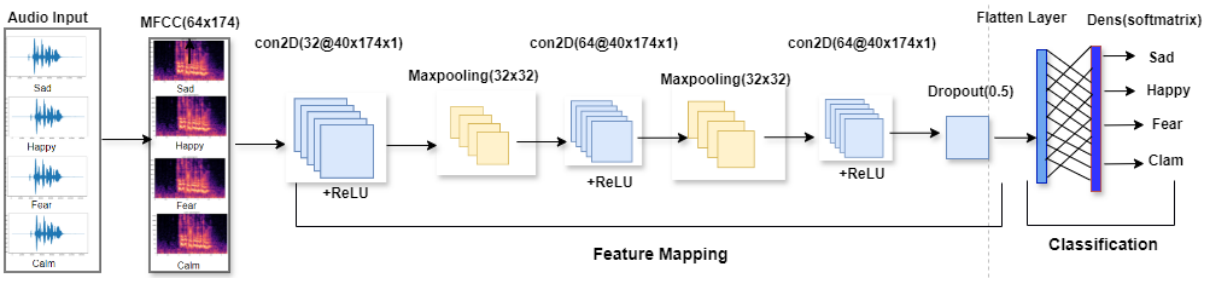

Figure 6: 2D-CNN Architecture

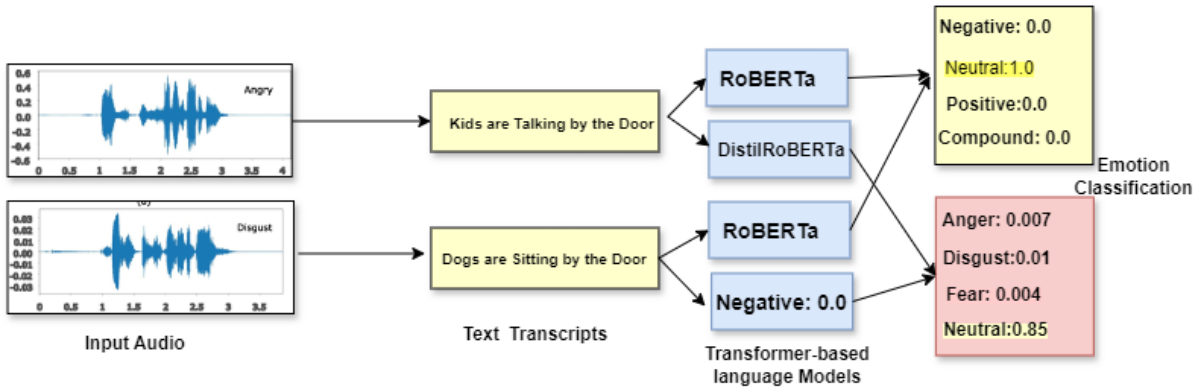

Figure 7: Text Sentiment Analysis using RoBERTa and DistilRoBERTa

The model development process for the multimodal sentiment analysis system involved sequential training of audio and text models. Initially, key features were extracted from the audio data, including Mel-Frequency Cepstral Coefficients (MFCCs), Chroma features, and Zero-Crossing Rate (ZCR), which captured pitch, timbre, and vocal intensity cues related to emotion. These features were preprocessed and normalized, with Mel spectrograms converted into image-like arrays suitable for convolutional input. Simultaneously, audio was transcribed using AssemblyAI's speech-to-text API, producing clean, tokenized text for sentiment classification. For model training, the dataset was split into training and testing sets using an 80:20 ratio. Label encoding was applied to categorical emotion labels, and Synthetic Minority Oversampling Technique (SMOTE) was used to balance the training data and mitigate class imbalance. Several models were explored for audio-based emotion recognition. A Random Forest classifier was implemented as a baseline using flattened MFCC features. Subsequently, a 1D Convolutional Neural Network (1D-CNN) was trained on sequential MFCC inputs to capture temporal patterns, while a more advanced 2D Convolutional Neural Network (2D-CNN) was trained on spectrogram images. The 2D-CNN architecture included convolutional, max-pooling, batch normalization, dropout, and dense layers, using ReLU as the activation function and softmax for multi class output. It was optimized using the RMSprop algorithm with a learning rate of 0.00005 and trained over multiple epochs with early stopping to prevent overfitting. Bidirectional LSTM as also implemented with layer of 128 units, designed to capture temporal dependencies in both forward and backward directions from input sequences shaped (174, 59). A dropout of 0.5 is applied for regularization to prevent overfitting. Following this, a dense layer with 128 units and a LeakyReLU activation (with alpha=0.1) adds non-linearity while maintaining gradient flow for negative inputs.

Another dropout layer with 0.5 rate is added for further regularization. Finally, the output layer uses a softmax activation to classify inputs into the number of emotion classes defined by *le.classes*. The model is compiled using the Adam optimizer with a learning rate of 0.001, sparse categorical cross-entropy loss, and accuracy as the evaluation metric.When the LSTM model was initially applied, it struggled to generalize well to the test data, achieving only about 55% accuracy, while the training accuracy was significantly higher at 81%. This discrepancy indicated overfitting, where the model learned the training features too specifically but failed to capture the underlying patterns needed for accurate predictions on unseen data. To address this, regularization techniques such as dropout were introduced, and model architecture adjustments were considered to improve generalization and reduce overfitting. For the text sentiment analysis component, a pre-trained DistilRoBERTa model from Hugging Face was employed due to its efficient transformer architecture and strong contextual understanding. The transcribed text was tokenized and fed into DistilRoBERTa to classify sentiments into Positive, Neutral, or Negative categories. Subsequently, the audio and text modalities were integrated, with predictions from both analyzed in parallel. Model performance was rigorously evaluated using metrics such as accuracy, precision, recall, F1-score, and confusion matrices to ensure reliable classification across multiple emotion classes. The best-performing 2D-CNN model achieved an accuracy of 72% on the RAVDESS dataset and 63% on EMOGATOR. All trained models were saved using joblib and seamlessly integrated into a Streamlit-based web interface. This platform supports real-time audio uploads, transcription via AssemblyAI, and simultaneous emotion detection from both speech and text inputs. This end-to-end development pipeline guarantees that the system operates with high accuracy and low latency, making it suitable for real-world applications.

5.3 Design Rationale

The design rationale behind the multimodal sentiment analysis system was to develop a robust and scalable framework that could detect emotional cues from both audio and text inputs with high accuracy and low latency. Key features such as MFCCs, Chroma, and ZCR were extracted to represent tonal and intensity-based emotional characteristics in speech, while Mel spectrograms enabled image-based pattern recognition through CNNs. Transcribed text from AssemblyAI was analyzed using DistilRoBERTa for its efficient transformer architecture and strong contextual language modeling. To mitigate class imbalance, SMOTE was employed, and categorical labels were encoded for consistent classification. Audio modeling explored various architectures—Random Forests as a baseline, 1D-CNN for temporal MFCC patterns, 2D-CNN for spatial spectrogram analysis, and a regularized BiLSTM with LeakyReLU activations to address overfitting. Predictions from both modalities were evaluated independently and then integrated for improved performance. The system's effectiveness was validated using accuracy, precision, recall, F1-score, and confusion matrices across RAVDESS and EMOGATOR datasets, achieving up to 72% accuracy. All models were saved with joblib and deployed via a Streamlit interface, creating a seamless pipeline for real-time sentiment detection that balances technical efficiency with practical usability.

6 Evaluation

6.1 Experiment /Audio Sentiment Analysis

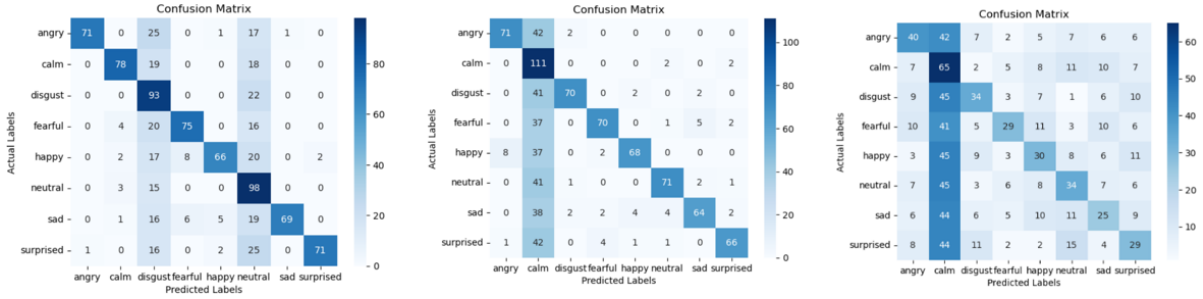

Figure 8: Random Forest - Mfcc,Mel Spectrogram,Chroma-RAVDESS

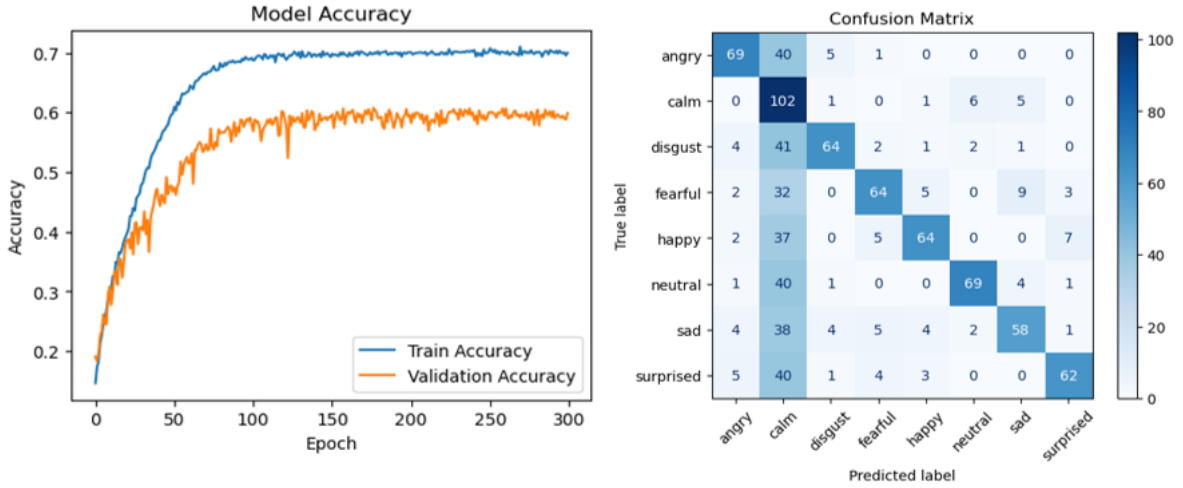

Figure 9: 2DCNN- Chroma-RAVDESS

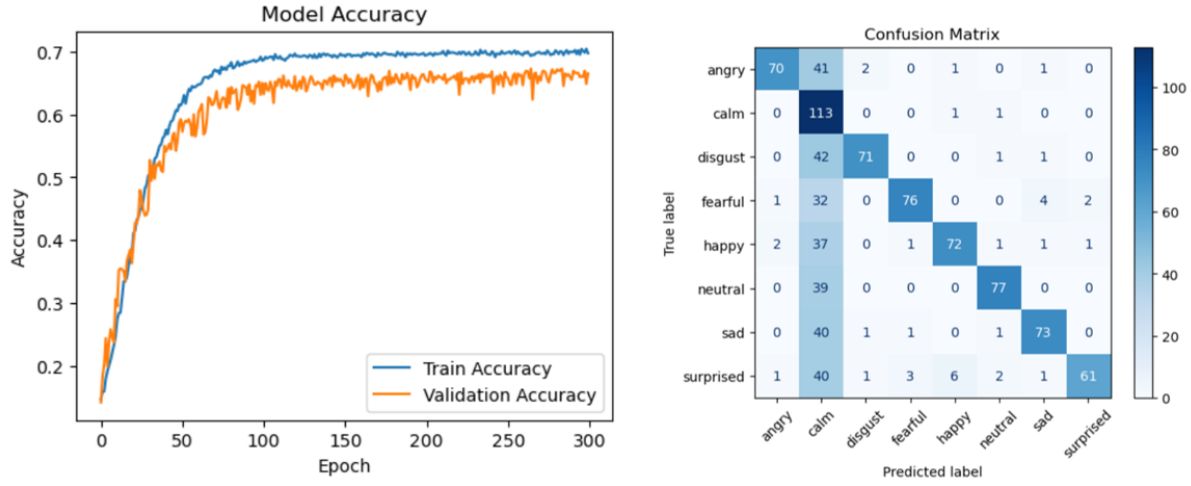

Figure 10: 2DCNN- Mel Spectrogram-RAVDESS

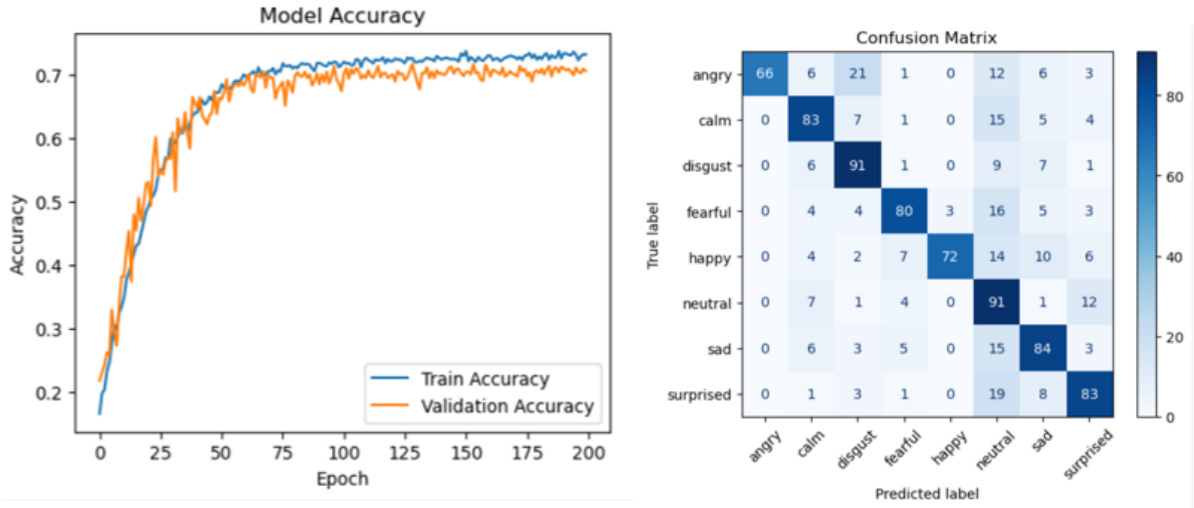

Figure 11: 2DCNN- MFCC-RAVDESS

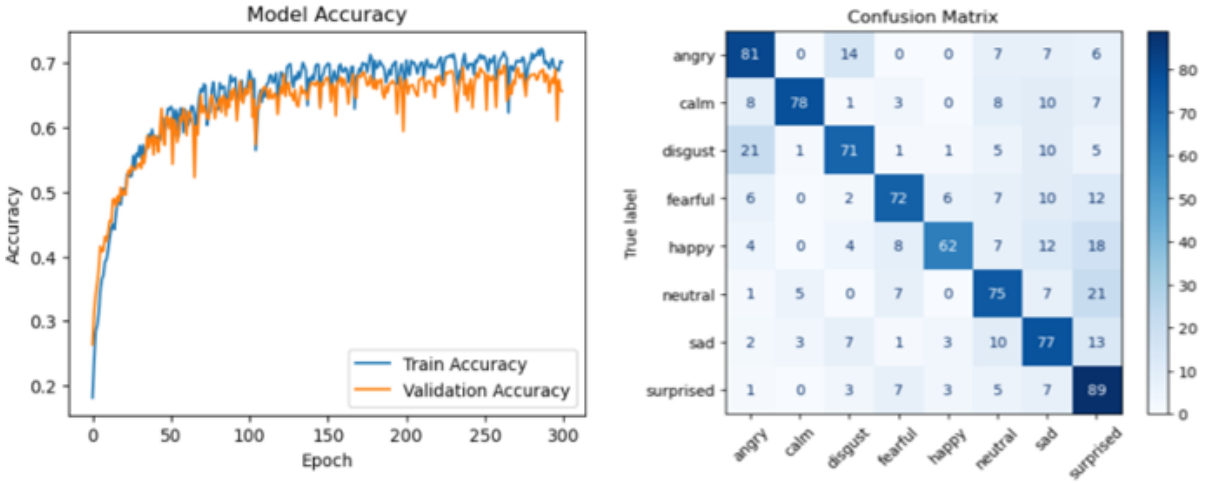

Figure 12: 2DCNN- biLSTM-RAVDESS

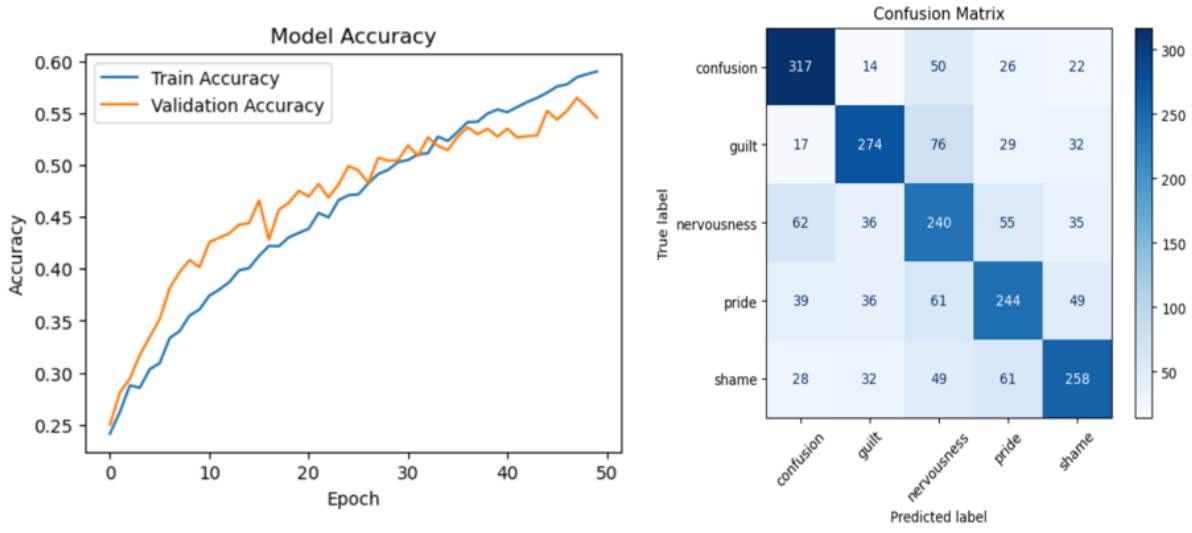

Figure 13: 2DCNN- EMOGATOR(MFCC)

6.2 Experiment /Text Sentiment Analysis

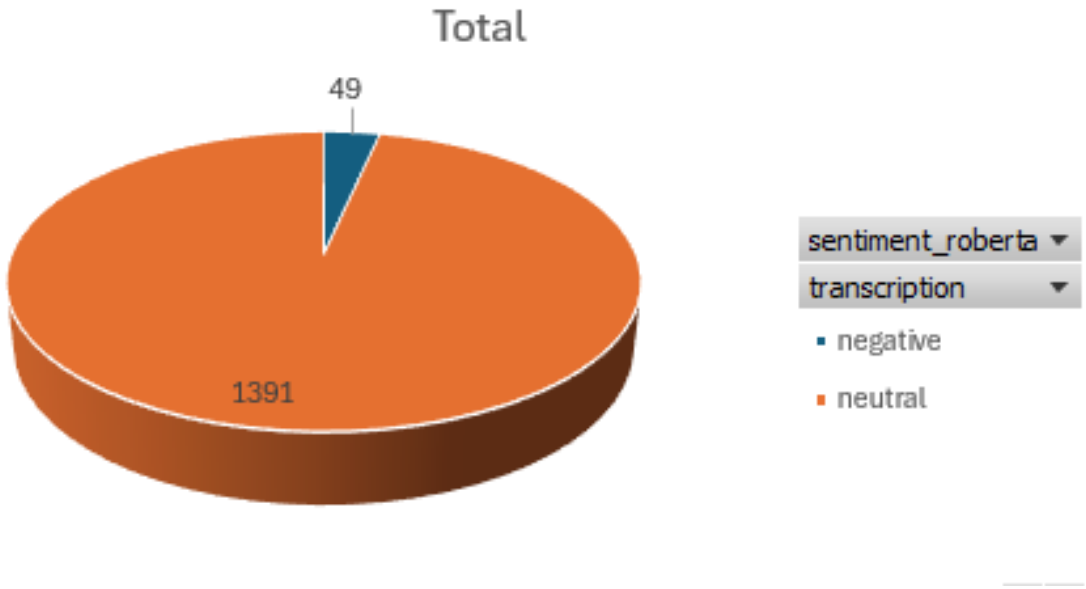

Figure 14: RoBERTa- Base Model - Text Sentiment Analysis

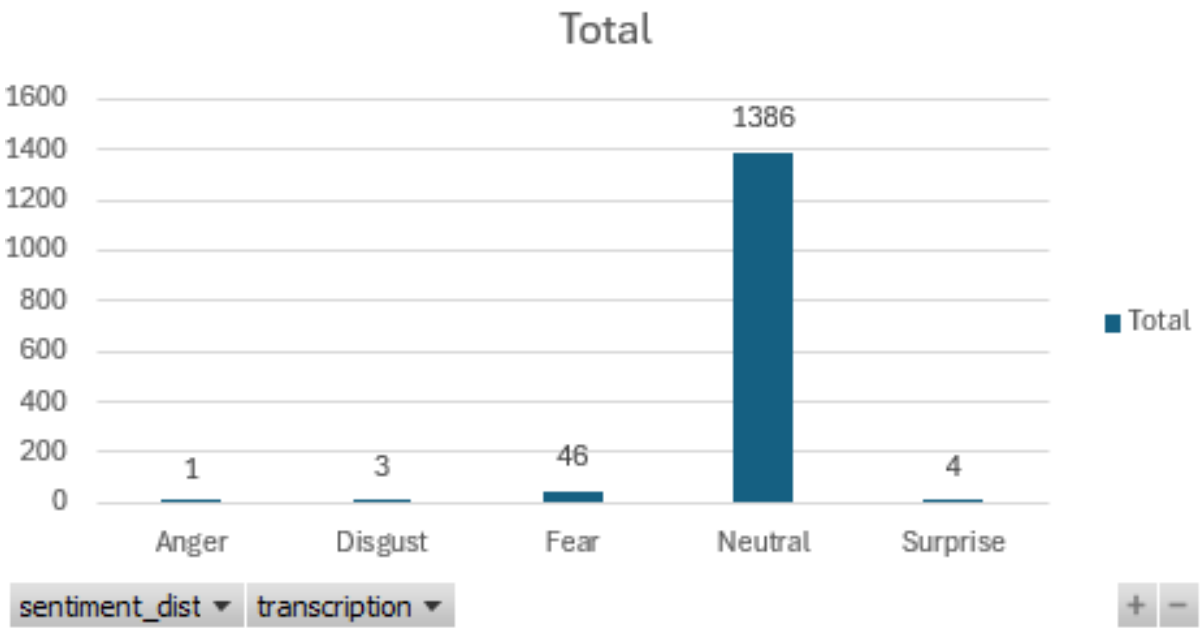

Figure 15: DistilRoBERTa- Base Model- Text Sentiment Analysis

6.3 Model Summary

Table 1: Audio Sentiment Analysis Results

Dataset	**Model**	**Features**	**Accuracy(%)**	**Training Time(s)**
RAVDESS (Original)	Random Forest	Mel,MFCC,Chroma	57,34,65.12,26.24	1.5s
RAVDESS (Augmented)	CNN (1D)	Mel,MFCC,Chroma	66.9,67.12,31.12	15m(avg)
RAVDESS	CNN (1D)	$Combined(Mel + MFCC + Chroma)$	67.21	65m
RAVDESS	CNN (2D)	Mel, **MFCC**, Chroma	67.12,**72.12**,60	15m(avg)
RAVDESS	CNN (2D)	$Combined(Mel + MFCC + Chroma)$	69.12	42m,10s
RAVDESS	CNN + biLSTM	$Combined(Mel + MFCC + Chroma)$	65.6	70m, 43s
EMOGATOR	Random Forest	Mel, MFCC, Chroma	57.2,61.12,47.28	2.4s
	CNN (1D)	Mel,MFCC,Chroma	53.2,63,45.28	17m(avg)
	CNN (2D)	Mel, **MFCC**, Chroma	57.2,**63**,47.28	17m(avg)
	CNN (2D)	$Combined(Mel + MFCC + Chroma)$	60.47	90m,2s
	CNN + biLSTM	$Combined(Mel + MFCC + Chroma)$	60.57	76m,2s

Table 2: F1 Scores for - minority emotions like disgust or calm

Model	**Emotion**	**Class**	**Precision**	**Recall**	**F1-Score**
2*BiLSTM - Combined Features	Calm	1	0.90	0.68	0.68
	Disgust	2	0.70	0.68	0.77
2*CNN(2D) + MFCC	Calm	1	0.71	0.72	0.72
	Disgust	2	0.69	0.79	0.74

Multimodal sentiment analysis combining audio and text yields stronger predictions than unimodal approaches. As highlighted in Table 1, 2D-CNN trained on spectrograms reached 72% accuracy on RAVDESS, excelling in emotion-rich acoustic capture. DistilRoBERTa effectively extracted nuanced textual sentiment but faltered with neutral tone ambiguity. Modality integration proved complementary—tone, pitch, and intensity enhanced classification synergy. Early LSTM overfitting highlighted temporal generalization challenges, mitigated via dropout and architecture tweaks. Transcription errors and class imbalance remain notable limitations needing refinement. Future work may explore audio transformers or attention-based fusion in noisy, diverse settings. Resampling strategies revealed k-fast outperforming k-best on 'Fearful' and 'Angry' classes with longer training times. k-fast likely preserves expressive emotional patterns better

under high-arousal conditions. Increasing epochs to 200 boosted 'Neutral' class performance, aiding low-expressivity feature learning. These results show training duration and resampling tuning impact emotion-class accuracy in imbalanced datasets. Using EMOGATOR, 2D-CNN reached 62% accuracy on subset data, confirming CNNs' robustness in low-resource setups. BiLSTM overfitted, struggling to generalize. Fig 14 results from text sentiment analysis, with RoBERTa classifying 1,391 text transcriptions as Neutral and 49 towards negative out of total of 1440 file of base model.Approximately 96.6% of the text transcriptions were classified as Neutral by RoBERTa. That's a strong skew toward neutrality—possibly indicating limitations in sentiment detection. Fig 15 results DistilRoBERTa classified 1386 out of 1440 customer service texts as Neutral, making up over 96% of the dataset. Fear was the most frequent non-neutral emotion, appearing in 46 instances. Surprise, Disgust, and Anger were rare, with only 4, 3, and 1 cases, respectively.

As illustrated in Table 2, the CNN(2D) model using MFCC features (Fig. 11) demonstrates a more consistent and balanced performance across both emotion classes when compared to the BiLSTM model (Fig. 12). While BiLSTM achieves higher precision for the "Calm" emotion, it does this at the expense of recall—an important consideration in real-world scenarios where failing to detect emotional cues can lead to missed opportunities for empathetic engagement. For the "Disgust" emotion, BiLSTM records a slightly higher F1-score (0.77 vs. 0.74); however, CNN(2D)'s superior recall (0.79) suggests it is more sensitive and responsive to emotional variations. This makes CNN(2D) a more reliable choice for applications requiring robust emotion detection across diverse user interactions.

7 Conclusion and Future Work

The development of the multimodal sentiment analysis system demonstrated that combining audio and text inputs significantly enhances emotion classification accuracy, providing a robust framework for real-time, user-centric applications. Mel spectrograms processed through 2D-CNNs effectively extract intricate acoustic features, while DistilRoBERTa captures contextual sentiment from transcribed speech. To address dataset imbalance and overfitting—particularly in LSTM-based architectures—techniques like SMOTE and dropout regularization were employed with strong effect. Integrating multiple models via parallel processing and deploying the system through a Streamlit web app further enabled smooth, low-latency user interaction. This architecture presents an ideal pathway for integration with CRM systems, offering dynamic analysis of customer sentiment from both speech and text in real time, and paving the way for smarter, emotionally aware user engagement. The RAVDESS dataset features actors vocalizing just two matched statements—"Kids are talking by the door" and "Dogs are sitting by the door"—across various emotional tones. While this controlled setup is excellent for isolating vocal emotion cues, it lacks the linguistic diversity needed for training models that rely on textual features. By contrast, MELD offers a much richer textual landscape. It contains over 13,000 utterances from 1,400 dialogues, all drawn from natural conversations in the Friends TV series3. Each utterance is annotated with both emotion and sentiment labels, and includes audio, visual, and textual modalities, making it ideal for multimodal emotion recognition tasks. Challenging two papers that directly address customer call center audio analysis using sentiment detection Atmaja and Sasou (2022),Dennison (2024b) the

experiments conducted successfully achieved the objective of delivering an end-to-end audio and text-based sentiment analysis platform. This includes a Streamlit web application and a custom-trained 2D-CNN model that reached 72% accuracy after a 15-minute training period on 4,320 audio files from RAVDESS and 13,000 audio files from EMOGATOR, which yielded 64% accuracy. Model evaluation based on Figures 8 through 13 highlights the best-performing architectures for audio sentiment classification. Additionally, Figures 14 and 15 present results from text-based sentiment analysis, offering complementary insights. Testing the end-to-end pipeline with real-time YouTube recordings demonstrates strong potential for audio-based sentiment analysis to enhance understanding of customer emotions—especially in cases where text alone may be insufficient. The integration of AssemblyAI within the Streamlit web app provides a production-ready solution for deployment within Customer Relationship Management (CRM) systems.

References

Alhussein, G., Ziogas, I., Saleem, S. and Hadjileontiadis, L. J. (2025). Speech emotion recognition in conversations using artificial intelligence: a systematic review and meta-analysis, *Artificial Intelligence Review* **58**(7): 198.

Ali, H. M. U., Farooq, Q., Imran, A. and Hindi, K. E. (2025). A systematic literature review on sentiment analysis techniques, challenges, and future trends, *Knowledge and Information Systems* **67**: 3967–4034.

Atmaja, B. T. and Sasou, A. (2022). Sentiment analysis and emotion recognition from speech using universal speech representations, *Sensors* **22**(17): 6369.

Cai, L., Liu, H., Zhang, N. and Huang, J. (2025). Bimodal sentiment analysis based on a pre-trained model and masked attention fusion, *IEEE Transactions on Audio, Speech and Language Processing* **33**: 2139–2150.

Dennison, J. (2024a). Emotions: Functions and significance for attitudes, behaviour, and communication, *Migration Studies* **12**(1): 1–20.

Dennison, J. (2024b). Emotions: functions and significance for attitudes, behaviour, and communication, *Migration Studies* **12**(1): 1–20.

Fernandez, A. and Awinat, S. (2024). Multimodal sentiment analysis based on video and audio inputs, *15th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN)*, Vol. 251, pp. 41–48.

Galić, J. and Đorđe Grozdić (2023). Exploring the impact of data augmentation techniques on automatic speech recognition system development: A comparative study, *Advances in Electrical and Computer Engineering* **23**(3): 3–12.

Hemmatian, J., Hajizadeh, R. and Nazari, F. (2025). Addressing imbalanced data classification with cluster-based reduced noise smote, *PLOS ONE* **20**(2): e0317396.

Khan, P. A., Sumanth, T. and Vardhan, K. V. (2021). Audio sentiment analysis, *International Journal of Creative Research Thoughts (IJCRT)* **9**(5).

Krishna, E. S. P., Ramu, T. B., Chaitanya, R. K., Ram, M. S., Balayesu, N., Gandikota, H. P. and Jagadesh, B. N. (2025). Enhancing e-commerce recommendations with sentiment analysis using mla-edtcnet and collaborative filtering, *Scientific Reports* **15**(1): 6739.

Luitel, S., Liu, Y. and Anwar, M. (2024). Investigating fairness in machine learning-based audio sentiment analysis, *AI and Ethics* **5**: 1099–1108.

Parmanand, S. P. and Anjali, D. (2025). Use of generative ai for audio speech recognition, *Proceedings of the International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)*, pp. 1113–1120.

Pathak, A. and Rathor, S. (2025). Abusive content detection with langchain integration using fine-tuned roberta, *2025 IEEE 14th International Conference on Communication Systems and Network Technologies (CSNT)*, pp. 1–5.

Płaza, M., Kazała, R., Koruba, Z., Kozłowski, M., Lucińska, M., Sitek, K. and Spyrka, J. (2022). Emotion recognition method for call/contact centre systems, *Applied Sciences* **12**(21): 10951.

Song, Y. and Zhou, Q. (2024). Bi-modal bi-task emotion recognition based on transformer architecture, *Applied Artificial Intelligence* **38**(1): 1–35.

Sutedja, I. and Hendry (2025). Sentiment analysis: An insightful literature review, *International Journal of Advanced Computer Science and Applications (IJACSA)* **16**(3).

Waleed, G. T. and Shaker, S. H. (2025). Speech emotion recognition on meld and ravdess datasets using cnn, *Information* **16**(7): 518.

Wei, Z., Ge, C. and Su, C. (2025). A deep learning model for speech emotion recognition on ravdess dataset, *International Journal of Advanced Computer Science and Applications* **16**(5): 252–258.

Yi, Y., Zhou, Y., Wang, T. and Zhou, J. (2025). Advances in video emotion recognition: Challenges and trends, *Sensors* **25**(12): 3615.

Yurtay, Y., Demirci, H., Tiryaki, H. and Altun, T. (2024). Emotion recognition on call center voice data, *Applied Sciences* **14**(20): 9458.