Sentiment Analysis from Audio Recordings

Department of Computer Science, School of Systems and Technology SST,
University of Management and Technology
Punjab, Pakistan, Lahore, 546832
June 2023

Sentiment Analysis from Audio Recordings

Abdul Moez[1], Ammara Malik[1] and Syed Arsalan Askari Naqvi[1]

Contributing authors: F2020266109@umt.edu.pk; f2020266540@umt.edu.pk; F2020266116@umt.edu.pk;

Abstract

Sentiment analysis from audio recordings is a process that utilizes natural language processing techniques and machine learning algorithms to analyze the emotional tone expressed in spoken language. By extracting features such as intonation, pitch, and tempo, these algorithms infer the underlying sentiment of the speaker. This analysis has diverse applications in various industries, including customer service, social media monitoring, and the entertainment industry. In customer service, it helps monitor and improve customer satisfaction during phone calls. In social media monitoring, it enables the tracking of public sentiment and the identification of trends. In the entertainment industry, sentiment analysis from audio recordings aids in gauging audience reactions and informing decision-making for future productions. This abstract highlights the significance of sentiment analysis from audio recordings in gaining valuable insights, making informed decisions, and enhancing experiences across different domains.

Keywords: Machine learning, classification models comparison, audio sentiment analysis, reducing noise, CNN classification model.

Overleaf editable link: Click-Here.

1 Introduction

Sentiment analysis, also known as opinion mining, has been a prominent task in Natural Language Processing (NLP) for analyzing and classifying textual data based on sentiment or opinion. However, the growing availability of audio data and the need to extract sentiment from spoken language have led to the emergence of audio sentiment analysis as a distinct research field. Audio sentiment analysis involves applying NLP techniques and machine learning algorithms to analyze the emotional tone and sentiment expressed in spoken language within audio recordings.

The motivation behind audio sentiment analysis stems from the recognition that spoken language carries unique emotional cues, such as intonation, pitch, and tempo, which are not fully captured in traditional text-based sentiment analysis. By leveraging these acoustic features, audio sentiment analysis aims to provide a more comprehensive understanding of human sentiment, enabling a deeper analysis of emotions, attitudes, and opinions expressed in audio data.

The applications of audio sentiment analysis are diverse and impactful. In customer service, it enables businesses to monitor and assess customer sentiment and satisfaction during phone calls, leading to improved service delivery and customer experiences. In social media monitoring, audio sentiment analysis helps identify trends and sentiments around specific topics, enhancing the understanding of public opinion and sentiment on platforms like podcasts, radio shows, and video content. Moreover, in the entertainment

industry, audio sentiment analysis provides valuable insights into audience reactions to films, TV shows, and live performances, aiding in decision-making for future productions and marketing strategies.

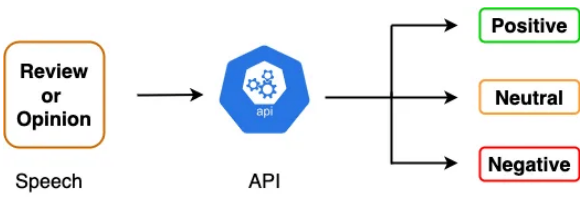

Fig. 1 An overview of the Sentiment Analysis model through an API .

However, audio sentiment analysis comes with its own set of challenges. The variability in speech patterns, accents, and background noise poses difficulties in accurately extracting sentiment from audio recordings. Additionally, the lack of large-scale labeled audio datasets adds complexity to the training of sentiment analysis models. Nonetheless, advancements in signal processing, acoustic feature extraction, and machine learning techniques have paved the way for significant progress in audio sentiment analysis.

In this project, we aim to explore and develop a robust audio sentiment analysis system by leveraging NLP techniques and machine learning algorithms. By analyzing acoustic features and patterns within audio recordings, we seek to accurately classify the emotional tone and sentiment expressed in spoken language. The outcomes of this project have the potential to benefit various industries, including customer service, social media monitoring, and the entertainment sector, providing valuable insights into human sentiment from audio data.

2 Motivation

The motivation for selecting this project lies in the need to capture and analyze the emotional cues present in spoken language, which are not adequately captured in text-based sentiment analysis. By leveraging acoustic features, such as intonation, pitch, and tempo, this project aims to provide a more comprehensive understanding of human sentiment expressed in audio recordings. This can lead to improved insights into emotions, attitudes, and opinions, enabling applications in customer service, social media monitoring, and the entertainment industry.

- **Capturing Rich Emotional Cues in Spoken Language:**
  Sentiment analysis from audio recordings offers a unique opportunity to tap into the rich emotional cues present in spoken language, providing a more comprehensive understanding of human sentiment compared to text-based analysis alone. By analyzing features, we can capture nuanced emotions that may be missed in written text, making it a compelling area of research.

- **Addressing the Importance of Voice-Based Platforms:**
  With the growing popularity of voice-based platforms and the increasing use of audio content in various domains, it becomes crucial for monitoring customer sentiment during phone calls, analyzing feedback in podcasts or radio shows, and assessing audience reactions to live performances or broadcasts.

- **Enhancing Customer Service Interactions:**
  The application in customer service can significantly enhance the quality of interactions by identifying customer frustration, satisfaction, or other emotions expressed during phone calls. This can lead to improved service delivery, customer retention, and overall customer satisfaction.

- **Gaining Deeper Insights in Social Media Monitoring:**
  Sentiment analysis from audio recordings can contribute to a more comprehensive understanding of public sentiment in social media monitoring. By capturing emotional nuances in audio content shared on platforms like YouTube or podcasts, we can gain deeper insights into public opinion, track trends, and respond effectively to emerging issues or concerns.

3 Methodology

Sentiment analysis uses machine learning models to perform audio analysis of human language. The working of audio sentiment analysis involves several steps that collectively enable the system to analyze and classify the emotional tone and sentiment expressed in audio recordings. Finally, data visualization is utilized to display the dataset trend and the final classification results. The details of each step are provided in the following subsection:

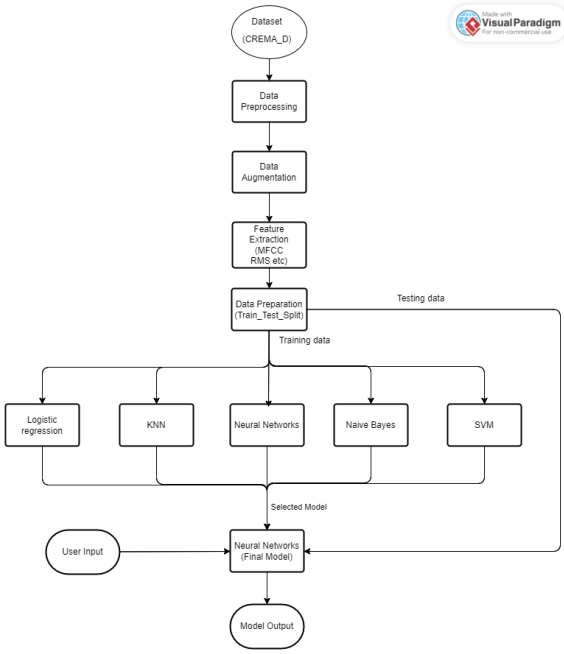

Fig. 2 The framework of the proposed prediction model for the sentiment analysis from audio recordings.

3.1 Data Collection

The first step in audio sentiment analysis is to gather a suitable dataset consisting of audio recordings that are labeled with sentiment classes, such as positive, negative, or neutral. This dataset can be collected from various sources, including customer service calls, audio reviews, social media platforms, or curated datasets.

We are using the CREMA-D dataset. [1] The CREMA-D dataset comprises 7,442 unique audio clips obtained from 91 actors, encompassing diverse demographics. These actors include 48 males and 43 females, spanning an age range of 20 to 74 and representing various races and ethnicities such as African American, Asian, Caucasian, Hispanic, and Unspecified. The dataset consists of recordings where the actors expressed 12 different sentences, each associated with one of six emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) and four emotion intensity levels (Low, Medium, High, and Unspecified).

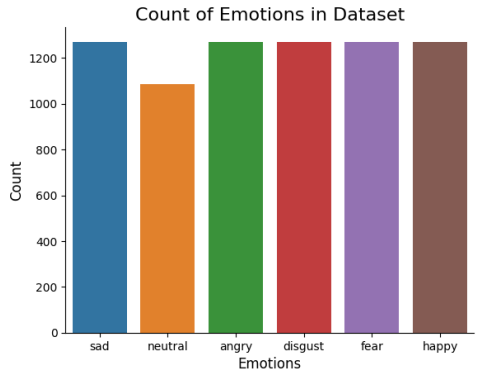

Fig. 3 Graph plot of the count of emotions in dataset.

3.2 Data Preprocessing:

Before performing sentiment analysis, the audio recordings need to be preprocessed. This includes steps such as audio normalization to adjust volume levels, noise removal to reduce background noise interference, and audio segmentation to divide long recordings into smaller, manageable segments. [?] Following are it's few steps:

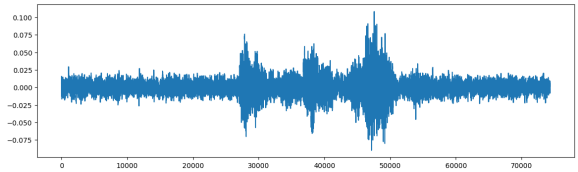

Fig. 4 Audio data with the noise.

- **Audio Normalization:**
  This is a crucial step in audio preprocessing that aims to adjust the volume levels of audio recordings. The main purpose of audio normalization is to ensure consistency and eliminate

variations in loudness across different audio files or segments within a file.

- **Noise Reduction:**
  Background noise can interfere with sentiment analysis. Noise reduction techniques are applied to minimize the impact of background noise and improve the quality of the audio signal. Noise reduction techniques contribute to improving the accuracy and reliability of sentiment analysis. By reducing background noise, the sentiment analysis model can focus more effectively on extracting relevant acoustic features and capturing the emotional tone expressed in the audio.

- **Feature Scaling:**
  In audio sentiment analysis, various acoustic features are extracted from the audio data, such as pitch, energy, and tempo. These features may have different scales and ranges. Feature scaling is performed to normalize the values of these features, ensuring they are on a similar scale. Common techniques for feature scaling include standardization (mean normalization) or normalization (min-max scaling).

- **Encoding Sentiment Labels:**
  The sentiment labels assigned to the audio segments, such as positive, negative, or neutral, need to be encoded into a numerical format for machine learning algorithms to process. This is typically done using one-hot encoding or label encoding. One-hot encoding represents each sentiment class as a binary vector, where only one element is "1" and the rest are "0". Label encoding assigns a unique numerical value to each sentiment class.

- **Handling Imbalanced Data:**
  In sentiment analysis, the distribution of sentiment classes in the dataset may be imbalanced, meaning some sentiment classes have significantly fewer samples than others. This can impact the model's performance. Preprocessing techniques like oversampling or undersampling can be applied to balance the dataset, ensuring equal representation of each sentiment class during model training.

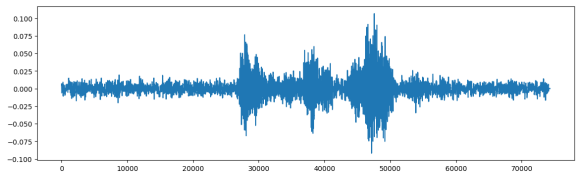

Fig. 5 Audio data after filtering the noise.

3.3 Data augmentation

Data augmentation is the process by which we create new synthetic data samples by adding small perturbations to our initial training set. [2] To generate syntactic data for audio, we can apply noise injection, shifting time, and changing pitch and speed. The objective is to make our model invariant to those perturbations and enhance its ability to generalize. In order for this to work, adding the perturbations must conserve the same label as the original training sample. In images data augmentation can be performed by shifting the image, zooming, rotating from the above types of augmentation techniques we are using noise, stretching(ie. changing speed), and some pitching.

3.4 Feature Extraction:

The next step involves extracting relevant acoustic features from the preprocessed audio segments. The audio signal is a three-dimensional signal in which three axes represent time, amplitude, and frequency. These features capture the characteristics of the speech signal that are indicative of sentiments, such as pitch, intonation, energy, rhythm, and tempo. Common techniques for feature extraction in audio sentiment analysis include Chroma-stft, Mel-frequency cepstral coefficients (MFCCs), prosodic features, and spectral features.

3.5 Model Training:

With the labeled audio segments and their corresponding acoustic features, a machine learning model is trained to classify sentiment. Various supervised learning algorithms are utilized, including logistic regression, support vector machines (SVM), K-nearest neighbor (KNN), Naive bayes, and convolutional neural networks

(CNNs). The training data is split into training and validation sets to evaluate the model's performance and fine-tune hyperparameters.

3.6 Model Evaluation:

The trained sentiment analysis model is evaluated using appropriate evaluation metrics, such as accuracy, precision, recall, and F1 score. Cross-validation techniques can be employed to assess the model's robustness and generalization capabilities. Additionally, confusion matrices and visualizations can provide insights into the model's performance across different sentiment classes.

3.7 Deployment and Testing:

Once the sentiment analysis model is deemed satisfactory in terms of performance, it can be deployed for testing on new, unseen audio recordings. The model predicts the sentiment class for each audio segment, providing insights into the emotional tone expressed in the audio. Post-processing techniques can be applied to further refine the sentiment analysis results, such as smoothing or filtering.

4 Data Visualization in audio sentiment analysis

Data visualization plays a significant role in audio sentiment analysis as it enables researchers and practitioners to gain insights and interpret the results effectively. Here are some key points regarding data visualization in audio sentiment analysis:

4.1 Acoustic Features Visualization:

Audio sentiment analysis often involves extracting various acoustic features, such as pitch, energy, and rhythm, from the audio data. Visualizing these features can provide a comprehensive understanding of how they vary across different sentiment classes. Plots, such as line graphs or scatter plots, can be used to visualize the distribution and patterns of these features.

4.2 Emotion-related Visualizations:

Sentiment analysis in audio recordings often involves the classification of emotions [3], such as happiness, anger, or sadness. Data visualization techniques, such as emotion heatmaps or emotion clouds, can be utilized to represent the prevalence and intensity of different emotions over time or across different segments of the audio.

4.3 Data Visualization Techniques:

Data visualization in audio sentiment analysis can be effectively achieved using spectrograms and wavegraphs. These visualizations provide valuable insights into the acoustic characteristics and sentiment patterns present in the audio data. Here's an explanation of how spectrograms and wavegraphs can be utilized for data visualization:

- **Spectrograms:**A spectrogram is a 2D representation that illustrates the frequency content of an audio signal over time. It displays the intensity of different frequencies present in the audio waveform. In sentiment analysis, spectrograms can be used to visualize the distribution of acoustic features related to sentiment, such as pitch, energy, and formant frequencies. By analyzing the spectrogram, one can observe changes in frequency characteristics across different sentiment categories, revealing distinct patterns and trends.

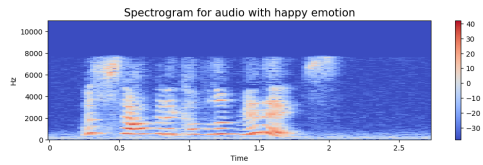

Fig. 6 Spectogram for audio with happy emotion

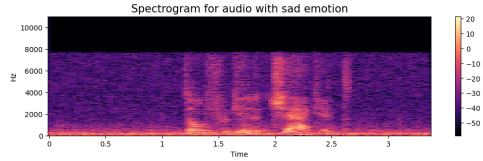

Fig. 7 Spectogram for audio with sad emotion

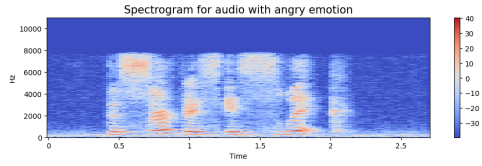

Fig. 8 Spectogram for audio with angry emotion

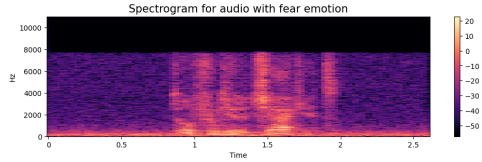

Fig. 9 Spectogram for audio with fear emotion

- **Waveplots:**A waveplot, also known as a waveform or audio waveform, represents the amplitude of the audio signal over time. It displays the waveform as a continuous plot, showcasing the variations in sound intensity. Waveplots are useful for visualizing the overall structure and temporal dynamics of the audio data. In sentiment analysis, Waveplots can provide insights into the intensity and duration of emotional expressions, allowing for the identification of specific moments or segments associated with different sentiment categories.

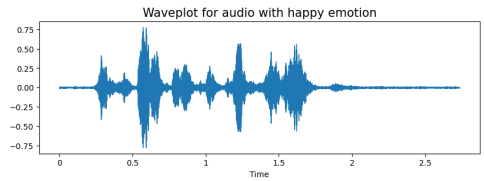

Fig. 10 Waveplot for audio with fear emotion

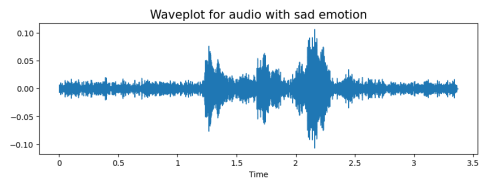

Fig. 11 Waveplot for audio with fear emotion

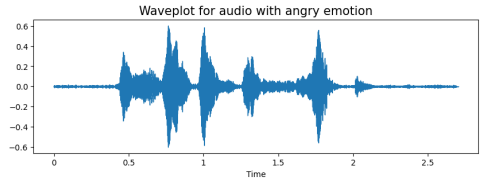

Fig. 12 Waveplot for audio with fear emotion

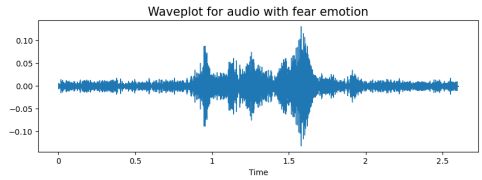

Fig. 13 Waveplot for audio with fear emotion

5 Performance Evaluation

In this phase of the project, the performance of different classification algorithms is evaluated based on accuracy. The accuracy results for each algorithm are provided, including Logistic Regression, KNN, SVM, Naive Bayes, and Neural Networks.

The performance of these classification models is assessed using various evaluation metrics, accuracy, precision, recall, and F1 score. The evaluation involves training the models on labeled audio data, validating them on a separate validation set, and testing their performance on an independent test set. The models are compared based on their ability to correctly classify audio recordings into the respective sentiment classes, and their overall performance is measured using the aforementioned evaluation metrics. [4]

5.1 Logistic Regression:

Logistic Regression is a binary classification algorithm that is commonly used in sentiment analysis. It models the relationship between the independent variables (acoustic features) and the binary sentiment class labels using a logistic function. It is efficient and interpretable, making it suitable for analyzing audio sentiment data.

Table 1 Classification Report - Logistic Regression.

	precision	recall	f1-score	support
angry	0.49	0.63	0.55	257
disgust	0.17	0.08	0.11	275
fear	0.15	0.13	0.14	251
happy	0.25	0.32	0.28	238
neutral	0.28	0.02	0.04	212
sad	0.39	0.72	0.50	256
accuracy			0.32	1489
macro avg	0.29	0.32	0.27	1489
weighted avg	0.29	0.32	0.28	1489

Table 2 Classification Report - K-Nearest Neighbors.

	precision	recall	f1-score	support
angry	0.50	0.56	0.53	257
disgust	0.24	0.29	0.26	275
fear	0.26	0.19	0.22	251
happy	0.32	0.31	0.31	238
neutral	0.22	0.24	0.23	212
sad	0.45	0.40	0.43	256
accuracy			0.33	1489
macro avg	0.33	0.33	0.33	1489
weighted avg	0.33	0.33	0.33	1489

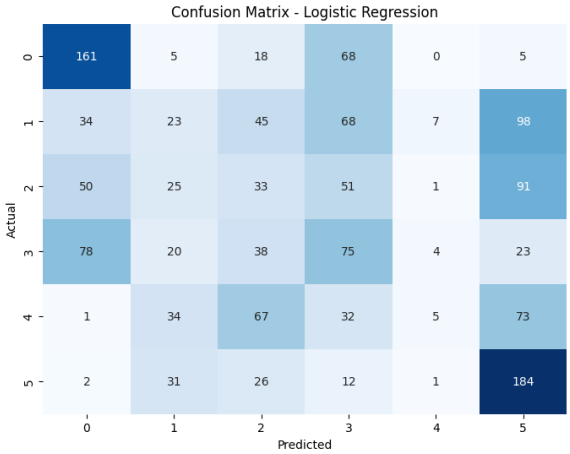

Fig. 14 Confusion matrix for Logistic Regression.

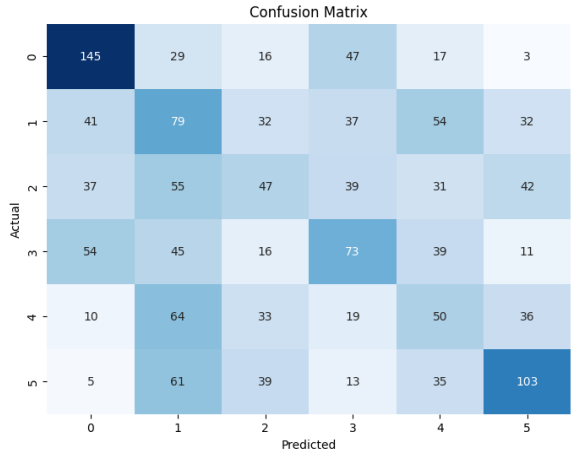

Fig. 15 Confusion matrix for the k-nearest neighbor (KNN).

5.2 K-Nearest Neighbors (KNN):

KNN is a non-parametric classification algorithm that assigns a data point to a class based on the classes of its nearest neighbors in the feature space. In audio sentiment analysis, KNN can be used to determine the sentiment class of an audio recording by considering the classes of similar audio samples in terms of their acoustic features.

5.3 Support Vector Machines (SVM):

SVM is a powerful classification algorithm that separates classes by finding an optimal hyperplane in the feature space. SVM seeks to maximize the margin between classes, allowing for effective sentiment classification. It can handle both linear and non-linear classification tasks and is widely used in sentiment analysis.

Table 3 Classification Report - Support Vector Machines.

	precision	recall	f1-score	support
angry	0.56	0.60	0.58	257
disgust	0.29	0.31	0.30	275
fear	0.25	0.10	0.14	251
happy	0.35	0.36	0.35	238
neutral	0.36	0.08	0.12	212
sad	0.35	0.73	0.48	256
accuracy			0.37	1489
macro avg	0.36	0.36	0.33	1489
weighted avg	0.36	0.37	0.33	1489

5.4 Naive Bayes:

Naive Bayes is a probabilistic classifier based on Bayes' theorem. It assumes independence among the features and calculates the posterior probability of a class given the observed features. Naive Bayes is known for its simplicity and efficiency,

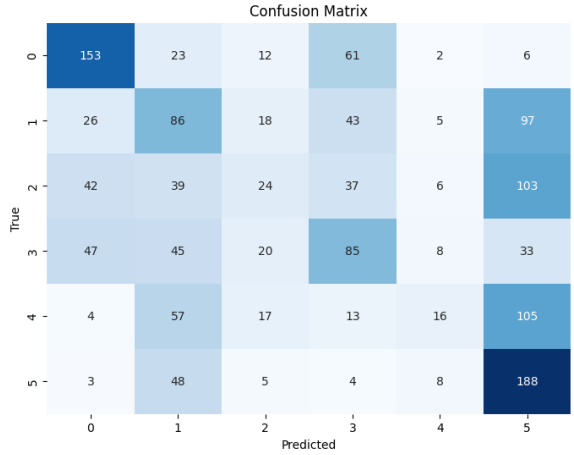

Fig. 16 Confusion matrix for support vector machines (SVM).

Table 4 Classification Report - Naive Bayes.

	precision	recall	f1-score	support
angry	0.63	0.44	0.52	257
disgust	0.21	0.16	0.18	275
fear	0.26	0.03	0.06	251
happy	0.27	0.16	0.21	238
neutral	0.20	0.26	0.23	212
sad	0.35	0.87	0.50	256
accuracy			0.33	1489
macro avg	0.32	0.32	0.28	1489
weighted avg	0.32	0.33	0.28	1489

making it suitable for sentiment analysis tasks with audio data.

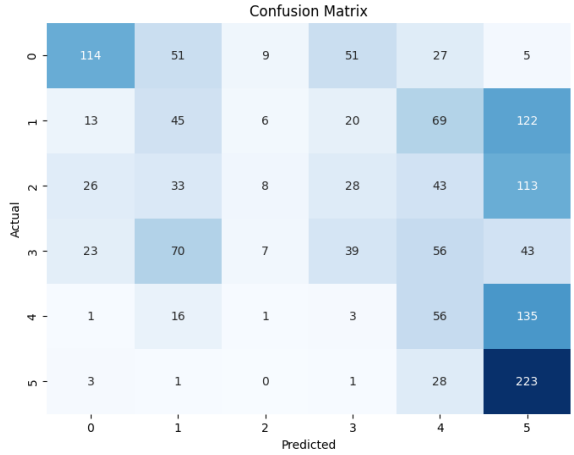

Fig. 17 Confusion matrix for the Naive Bayes algorithm.

Table 5 Classification Report - Neural Networks (CNN).

	precision	recall	f1-score	support
angry	0.57	0.61	0.59	957
disgust	0.33	0.36	0.34	912
fear	0.40	0.27	0.32	930
happy	0.39	0.34	0.36	999
neutral	0.36	0.45	0.40	786
sad	0.52	0.57	0.54	999
accuracy			0.43	5583
macro avg	0.43	0.43	0.43	5583
weighted avg	0.43	0.43	0.43	5583

5.5 Neural Networks:

Neural Networks are a class of models inspired by the functioning of the human brain. They consist of interconnected layers of nodes (neurons) that process input data. In audio sentiment analysis, neural networks can learn complex patterns and relationships between acoustic features and sentiment labels. They have achieved state-of-the-art performance in various NLP tasks, including sentiment analysis

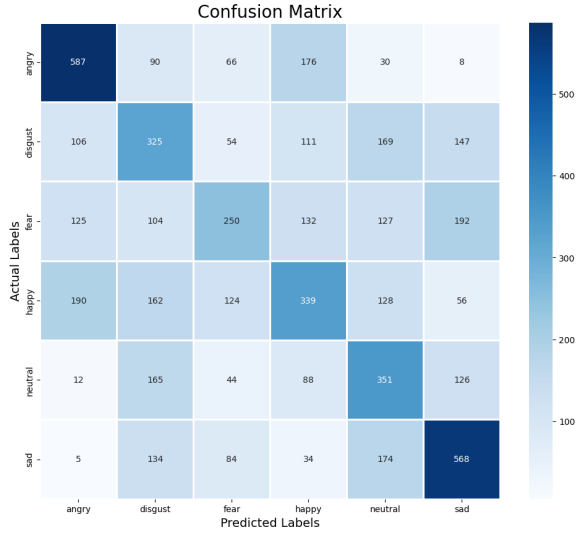

Fig. 18 Confusion matrix for the Neural Networks (CNN).

By evaluating the performance of these models, we can determine that Neural Networks perform best for audio sentiment analysis tasks, providing insights into the suitability and effectiveness of different approaches in capturing and

analyzing sentiment from audio data. Therefore, we used it as our model.

6 Predicted Result

After selecting a Convolutional Neural Network (CNN) as the model for audio sentiment analysis, we applied it to predict the sentiment labels of the test data. [5] The CNN model, with its ability to capture spatial and temporal dependencies in audio data, has shown promising performance in previous experiments.

Upon evaluating the test data using the trained model, we have observed encouraging results. The model has successfully classified the audio samples into their respective sentiment classes, demonstrating its effectiveness in capturing and understanding the emotional cues present in the audio recordings. The predicted sentiment labels provide insights into the sentiment expressed in the test audio data, enabling a deeper understanding of the emotional tone and attitude conveyed.

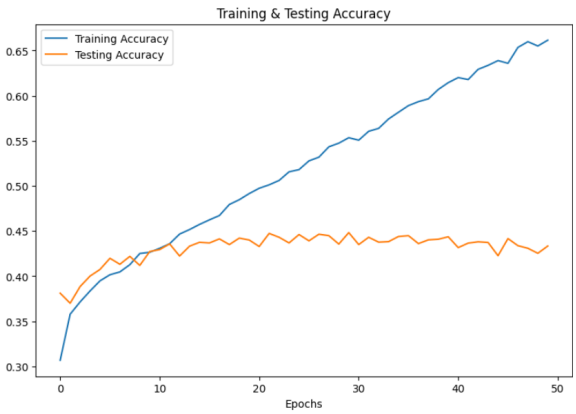

Fig. 19 Graph plot of accuracy in training and testing of the model.

The accuracy and performance metrics of the model on the test data demonstrate its robustness and reliability in audio sentiment analysis. The model's ability to effectively analyze and classify sentiment in audio recordings opens up possibilities for various applications, including customer service analysis, social media monitoring, and entertainment industry feedback.

Overall, the utilization of CNN as the chosen model in audio sentiment analysis has yielded promising results, showcasing its efficacy in accurately predicting sentiment labels for unseen audio data. This success highlights the potential of CNNs in capturing and analyzing emotional cues in audio recordings, paving the way for further advancements and applications in this field.

Table 6 Predicted results after applying the model.

	Predicted Lables	**Actual Lables**
0	fear	sad
1	neutral	sad
2	neutral	sad
3	disgust	angry
4	angry	angry
5	disgust	angry
6	sad	sad
7	sad	sad
8	neutral	sad
9	fear	fear

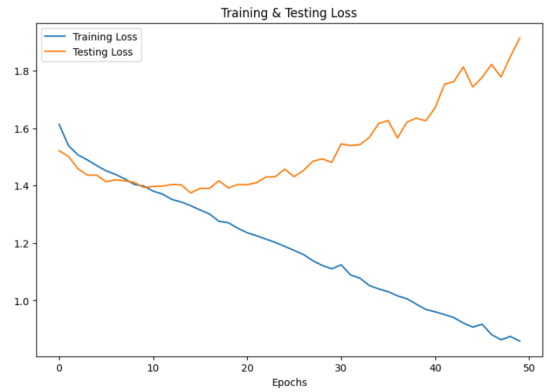

Fig. 20 Graph plot of loss in training and testing of the model.

7 Conclusion

In conclusion, the project on audio sentiment analysis has successfully addressed the need to capture and analyze emotional cues embedded in spoken language. By leveraging acoustic features such as intonation, pitch, and tempo, we have gained a more comprehensive understanding of human sentiment expressed in audio recordings.

Throughout this project, we have observed the potential of audio sentiment analysis in various domains. In customer service, it can serve as a powerful tool to monitor customer sentiment and satisfaction during phone calls, enabling companies to enhance their services and address

customer needs effectively. Additionally, in social media monitoring, audio sentiment analysis can help identify trends and sentiment around specific topics, enabling organizations to gain insights into public opinion and make informed decisions.

Looking ahead, further advancements in audio sentiment analysis hold the potential to improve the accuracy and reliability of sentiment classification. The development of more sophisticated machine learning algorithms and the availability of large-scale annotated audio datasets will contribute to the refinement of sentiment analysis models. This, in turn, will enable more precise monitoring of emotions, attitudes, and opinions expressed in audio recordings.

In conclusion, this project has shed light on the importance and potential applications of audio sentiment analysis. It serves as a stepping stone for further research and innovation in this field, leading to advancements in sentiment analysis techniques and the effective utilization of emotional cues in audio data for various practical purposes.

References

[1] M. R. Kabir, M. M. Muhaimin, M. A. Mahir, M. M. Nishat, F. Faisal, and N. N. I. Moubarak, "Procuring mfccs from crema-d dataset for sentiment analysis using deep learning models with hyperparameter tuning," in *2021 IEEE International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things (RAAICON)*. IEEE, 2021, pp. 50–55.

[2] R. Shankar, A. H. Kenfack, A. Somayazulu, and A. Venkataraman, "A comparative study of data augmentation techniques for deep learning based emotion recognition," *arXiv preprint arXiv:2211.05047*, 2022.

[3] S. K. Panda, A. K. Jena, M. R. Panda, and S. Panda, "Speech emotion recognition using multimodal feature fusion with machine learning approach," *Multimedia Tools and Applications*, pp. 1–19, 2023.

[4] E. Ghaleb, M. Popa, and S. Asteriadis, "Metric learning-based multimodal audio-visual emotion recognition," *Ieee Multimedia*, vol. 27, no. 1, pp. 37–48, 2019.

[5] M. Hosain, M. Arafat, G. Z. Islam, J. Uddin, M. M. Hossain, F. Alam *et al.*, "Emotional expression detection in spoken language employing machine learning algorithms," *arXiv preprint arXiv:2304.11040*, 2023.

Abdul Moez is currently a student of UMT doing his BS in Computer Science. His interests include Web Development, App development, Graphic Designing and problem solving. He plans to continue learning new technologies during his career in development and build an organization of his own.

Ammara Malik is currently a student of UMT studying computer science. She has always been fascinated by the world of technology and its endless possibilities. She is constantly seeking opportunities to improve her skills and knowledge.

Syed Arsalan Askari Naqvi is currently a student of UMT doing his BS in Computer Science. His interests include Web Development and drop shopping. He plans to continue learning new technologies during his career in development and build a business of his own.