Sentiment Analysis from Audio Recordings

Amaan Majeed[1*], Munib Ahsan Khan and Adam Saleem[†]

Department of Computer Science, School of Systems and Technology SST, University of Management and Technology, UMT Road, Lahore, 54000, Punjab, Pakistan.

*Corresponding author(s). E-mail(s): F2020266286@umt.edu.pk;
Contributing authors: F2020266270@umt.edu.pk; F2020266300@umt.edu.pk;
[†]These authors contributed equally to this work.

Abstract

Sentiment analysis, the process of extracting emotions and attitudes from various forms of data. It has gained a lot of significant attention in recent years. While sentiment analysis from textual data has been studied extensively, analyzing sentiments from audio recordings has emerged as an exciting research area. This approach opens up the possibility of analyzing the emotions of a person irrespective of what language they are speaking in. This research paper aims to address the challenges of sentiment analysis from audio recordings by employing machine learning algorithms. Specifically, it focuses on developing a robust and accurate model capable of classifying audio recordings into different emotion categories, including sadness, anger, disgust, fear, happiness, and neutral. The methodology encompasses several key steps, including Data Pre-processing, data augmentation, feature extraction, and the implementation and evaluation of various machine learning algorithms. The results and analysis section presents the performance of each algorithm, providing insights into their effectiveness in sentiment analysis from audio recordings. The findings contribute to the field of sentiment analysis and have the potential to benefit researchers and practitioners in related domains.

Overleaf editable link: Click-HERE.
Sentiment Analysis - CNN: Click-HERE.
Logistic Regression, KNN, NB, SVM: Click-HERE.

1 Introduction

Analyzing sentiments from audio recordings opens up new avenues for understanding human emotions in real-world scenarios. By capturing vocal cues, tone, and intonation, audio-based sentiment analysis provides a more comprehensive understanding of an individual's emotional state. This can be valuable in various domains, including customer feedback analysis, market research, mental health monitoring, and voice assistants that adapt their responses based on user emotions.

This research paper aims to contribute to the field of sentiment analysis by addressing the challenges of sentiment analysis from audio recordings. The objective is to develop a robust and accurate model capable of classifying audio recordings into different emotion categories, including sadness, anger, disgust, fear, happiness, and neutral. To achieve this goal, a comprehensive methodology encompassing several key steps is proposed.

The methodology begins with audio data preprocessing, where techniques such as noise reduction and audio file cleaning are applied to optimize

the input data for subsequent analysis. Data augmentation methods are explored to enhance the training dataset, introducing variations into the audio recordings to improve the model's ability to generalize and capture different variations of emotions.

Feature extraction plays a vital role in sentiment analysis. The paper presents a comprehensive analysis of various audio features employed in this study, including zero-crossing rate, chroma_stft, MFCC, RMS (root mean square) value, and MelSpectrogram[3]. These features capture distinct aspects of audio signals, providing valuable information for training the sentiment analysis model.

A machine learning pipeline is implemented, incorporating several algorithms such as logistic regression, K-nearest neighbors (KNN), support vector machines (SVM), naive Bayes, and convolutional neural networks (CNN). The performance of each algorithm is evaluated using metrics such as accuracy, precision, recall, and F1 score, providing insights into their effectiveness in accurately classifying audio recordings into different emotion categories.

The results and analysis section presents the performance of each algorithm and offers a comparative analysis of their effectiveness in sentiment analysis from audio recordings. The findings shed light on the strengths and limitations of each approach, providing insights into the suitability of different algorithms for this task.

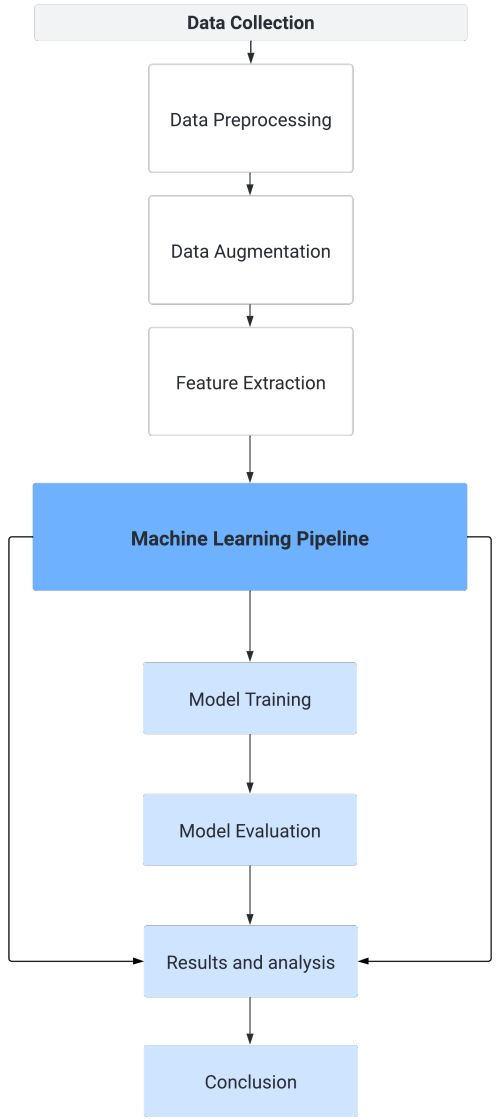
By exploring the use of various machine learning algorithms for sentiment analysis from audio recordings, this research paper aims to contribute to the field of sentiment analysis. The findings and insights obtained from this study have the potential to benefit researchers and practitioners working in areas such as natural language processing, affective computing, and human-computer interaction.

Overall, this research paper establishes a comprehensive methodology for sentiment analysis from audio recordings, combining preprocessing techniques, data augmentation, feature extraction, and machine learning algorithms. The subsequent sections provide detailed information on each step, including the methodologies employed and the rationale behind them.

The Machine learning algorithms used in this project are:

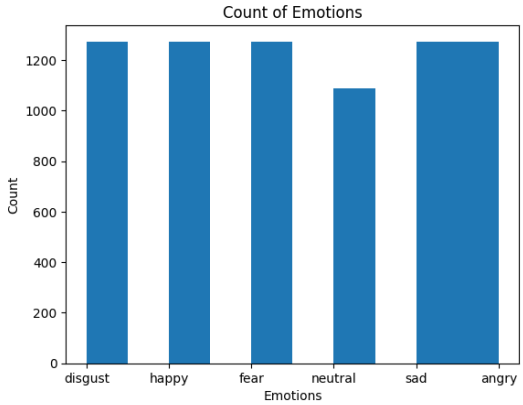
1. Logistic Regression
2. K-nearest neighbours(k-NN)
3. Naive Baye's
4. Support Vector Machines(SVM)

2 Methodology



2.1 Data Collection

A diverse dataset of audio recordings was collected, covering various speech styles and emotional states. The dataset originally belonged to a kaggle entry but it was including noise [5] in it. The dataset was uploaded to Google Drive and the "glob" function on python was used to read it from the Google Drive.

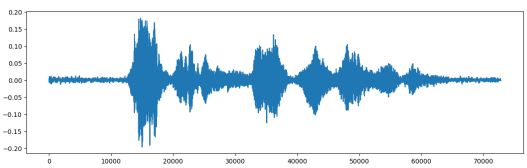


2.2 Preprocessing

The audio recordings were subjected to preprocessing steps to enhance the quality and facilitate feature extraction. This involved audio normalization, noise removal, and segmentation into shorter segments suitable for analysis.

2.3 Data Augmentation

Data augmentation is the process by which we create new synthetic data samples by adding small perturbations to our initial training set. We can apply noise injection, shifting time, and changing pitch and speed to generate syntactic data for audio. The objective is to make our model invariant to those perturbations and enhance its ability to generalize. In order to this to work adding the perturbations must conserve the same label as the original training sample.



2.4 Feature Extraction

The extraction of features is a crucial part of analyzing and finding relations between different things. As we already know that the data provided by audio cannot be understood by the models directly so we need to convert them into an understandable format for which feature extraction[2] is used. The audio signal is a three-dimensional signal in which three axes represent time, amplitude, and frequency.

2.5 Labeling

To train the sentiment analysis model, the audio segments were manually labeled[4] with sentiment labels[7] such as happy, sad, angry, fear, disgust or neutral. Human annotators listened to the audio and annotated the sentiment accordingly. Labels were extracted from the audio files then a dataframe called "Crema_df" was created to store all names of files and ylabels(ylabels are the emotions of audio files) and finally a waveplot and spectrogram were plotted for each emotion whereas only a waveplot was plotted for loudness of the audio.

2.6 Computation

After all the previous steps, the following for computed:

1. Zero Crossing Rate(ZCR):
   ZCR is the rate at which a signal changes from positive to zero to negative or from negative to zero to positive.

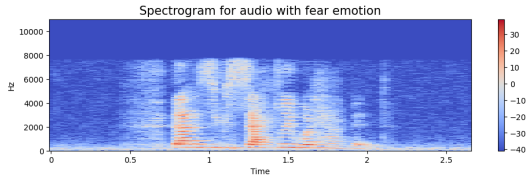
2. Chroma STFT:
   Short-term Fourier Transform represents information about the classification of pitch and signal structure. It depicts high values with a spike.

3. Mel Frequency Cepstral Coefficients(MFCC):
   MFCC[3] of a signal is a small set of features (usually about 10-20) that concisely describe the overall shape of a spectral envelope.

4. Root Mean Square Error (RMSE):
   RMSE is taking the square root of a mean squared error. It indicates the absolute fit of the model to the data and Provides average

model prediction error in units of the variable of interest.

5. Mel Spectogram:
   Mel Spectrogram applies a frequency-domain filter bank to audio signals that are windowed in time.



6. Chromagram: Chromagram is computed using the Short-Time Fourier Transform (STFT) and represents the distribution of energy across different pitch classes or musical notes.
7. Spectral Centroid: Spectral Centroid is a measure of the center of mass of the power spectrum and represents the average frequency content of the audio signal.
8. Spectral Bandwidth: Spectral Bandwidth is a measure of the spread of the power spectrum around the Spectral Centroid and provides information about the width of the frequency content in the signal.
9. Spectral Contrast: Spectral Contrast[1] measures the difference in amplitude between peaks and valleys in the power spectrum, providing information about the perceptual contrast of different frequency bands in the audio signal.

2.7 Model Training

Various Machine learning models such as Logistic Regression. k-NN, Naive Baye's and SVM (already mentioned before) were trained using labeled data. These models learned to associate audio features with sentiment labels.

3 Results

Table 1 Performance Metrics for Machine Learning Algorithms

Algorithm	Accuracy	Precision	Recall	F1 Score
Logistic Reg.	0.353	0.283	0.353	0.300
KNN	0.309	0.310	0.309	0.306
Naive Bayes	0.347	0.312	0.347	0.299
SVM	0.355	0.333	0.355	0.321
CNN	0.4317	0.4296	0.4317	0.4263

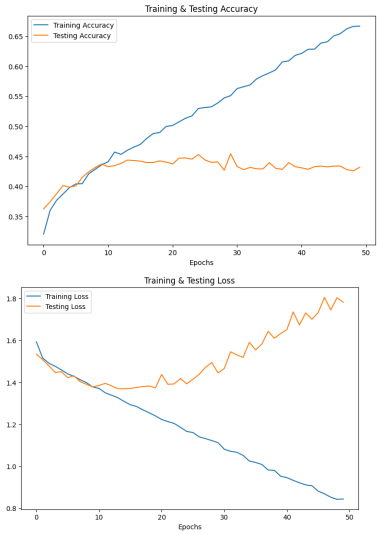
The performance of various machine learning algorithms was evaluated on the dataset using four evaluation metrics: accuracy, precision, recall, and F1 score. The results are summarized in Table 1.

Logistic Regression achieved an accuracy of 0.353, indicating that it correctly classified 35.3% of the samples. The precision score of 0.283 suggests that out of the samples predicted as positive, only 28.3% were truly positive. The recall score of 0.353 reflects the algorithm's ability to identify 35.3% of the actual positive samples. The F1 score of 0.300 indicates a balanced performance between precision and recall.

K-Nearest Neighbors (KNN) obtained an accuracy of 0.309, which implies a correct classification rate of 30.9%. The precision score of 0.310 indicates that approximately 31.0% of the samples predicted as positive were truly positive. The recall score of 0.309 suggests that the algorithm captured 30.9% of the actual positive samples. The F1 score of 0.306 represents the harmonic mean of precision and recall.

Naive Bayes achieved an accuracy of 0.347, implying a correct classification rate of 34.7%. The precision score of 0.312 suggests that approximately 31.2% of the samples predicted as positive were truly positive. The recall score of 0.347 indicates that the algorithm identified 34.7% of the actual positive samples. The F1 score of 0.299 represents the balanced performance between precision and recall.

Support Vector Machine (SVM) attained an accuracy of 0.355, signifying a correct classification rate of 35.5%. The precision score of 0.333 implies that approximately 33.3% of the samples predicted as positive were truly positive. The recall score of 0.355 reflects the algorithm's ability to capture 35.5% of the actual positive samples. The F1 score of 0.321 represents a balanced performance between precision and recall.



Overall, the results demonstrate the varying performance of different machine learning algorithms on the dataset. While Logistic Regression and SVM achieved relatively higher accuracy, the precision and recall scores provide insights into the algorithms' ability to correctly classify positive samples. KNN and Naive Bayes demonstrated comparable performance in terms of accuracy, precision, recall, and F1 score.

4 Applications

Sentiment analysis from audio recordings has various potential applications across different domains:

1. Customer Feedback Analysis:
   Sentiment analysis can be applied to customer service recordings, enabling companies to assess customer satisfaction levels and identify areas for improvement.
2. Market Research:
   Analyzing sentiment in recorded interviews, focus groups, or social media audio clips can provide valuable insights into consumer opinions, preferences, and trends.
3. Call Center Monitoring:
   Sentiment analysis can be used in real-time to monitor customer interactions, identify dissatisfied customers, and prompt proactive interventions.
4. Psychological Analysis: Sentiment analysis from therapy sessions or counseling recordings can assist psychologists in evaluating emotional states and tracking progress over time.

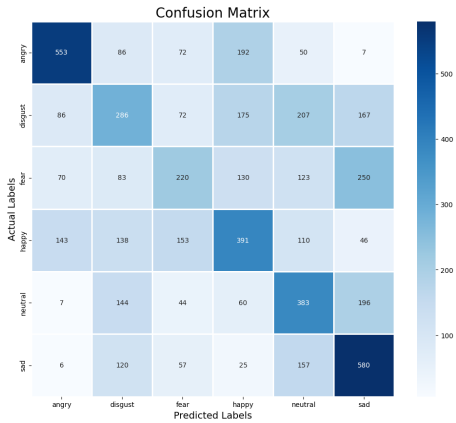
5 Limitations and Improvements

1. Multilingual Support:
   The current project focused on sentiment analysis in a single language. Extending the model to support multiple languages could enhance its usability and applicability.
2. Accent and Dialect Variations:
   The model's performance might be affected by variations in accents, dialects, and speech patterns. Fine-tuning the model on diverse regional data could help address this limitation.
3. Robustness to Noise:
   The model's performance might degrade in the presence of background noise or low-quality audio recordings. Further research on noise reduction techniques[6] and robust feature extraction could improve performance.
4. Real-time Analysis:
   Implementing the sentiment analysis model in real-time applications, such as live call monitoring, could enhance its usefulness in dynamic environments.

6 Conclusion

The research paper presents a comprehensive study on sentiment analysis from audio recordings, aiming to address the challenges and explore the potential of this domain. Through various stages of data collection, preprocessing, data augmentation, feature extraction, and machine learning, the paper develops a robust methodology for analyzing emotions and attitudes expressed in audio data.

The findings of this research indicate that sentiment analysis from audio recordings offers valuable insights into understanding human emotions in real-world scenarios. By capturing vocal cues, tone, and intonation, audio-based sentiment analysis provides a more comprehensive understanding of an individual's emotional state. This has significant implications for various domains, including customer feedback analysis, market research, mental health monitoring, and voice assistants that adapt their responses based on user emotions.



The evaluation of different machine learning algorithms, including logistic regression, K-nearest neighbors (KNN), naive Bayes, support vector machines (SVM), and convolutional neural networks (CNN), demonstrates their effectiveness in classifying audio recordings into different emotion categories. The results highlight the strengths and limitations of each approach and provide insights into the suitability of different algorithms for sentiment analysis from audio recordings.

The research paper contributes to the field of sentiment analysis by presenting a comprehensive methodology that encompasses various key steps. The data preprocessing techniques, data augmentation methods, and feature extraction approaches employed in this study contribute to optimizing the input data and capturing relevant information for sentiment analysis. The machine learning pipeline, incorporating multiple algorithms, offers flexibility and scalability in the analysis process.

Furthermore, this research paper opens up new avenues for future work in sentiment analysis from audio recordings. Additional research can focus on exploring more advanced machine learning techniques, such as deep learning models, to further enhance the accuracy and performance of sentiment analysis. Moreover, investigating the impact of different languages, cultural backgrounds, and demographic factors on audio-based sentiment analysis can provide a more nuanced understanding of emotions in diverse contexts.

In conclusion, sentiment analysis from audio recordings holds immense potential in understanding human emotions and attitudes. The research paper establishes a comprehensive methodology and presents valuable insights into the effectiveness of different machine-learning algorithms for this task. The findings and methodologies presented in this paper point to the CNN technique which showed the best result of all the other techniques. It contributes to the field of sentiment analysis and provides a foundation for further advancements in analyzing emotions from audio data.

7 Authors

- Amaan Majeed: Amaan is a 6th semester Student who is currently studying Machine learning as an elective. He is the president of the Computer Science Society in UMT 2022-2023 and was selected as a Google Developer Student Club's lead for the year. He is passionate about technology and is very keen on building things which create an impact.

- Munib Ahsan Khan: Munib is in the 6th Semester, having Machine learning as an elective. He is good with Databases and has a strong grip on his Algorithms. He is an expert in Game Designing, who is getting in touch with MindStrom media group for an internship

- Adam Saleem: Adam is doing his bachelor's from UMT. He is in the 6th semester and has a great sense of Web development. He has designed websites for clothing brands and is really looking towards expanding his portfolio on web development.

References

[1] Daniel PW Ellis. Mfccs and related spectral coefficients. In Daniel Presser W., editor, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*, pages 101–130. Wiley-IEEE Press, 2007.

[2] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[3] Beth Logan. Mel frequency cepstral coefficients for music modeling. In *International Symposium on Music Information Retrieval (ISMIR)*, 2000.

[4] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matthew McVicar, Eric Battenberg, ..., and Oriol Nieto. Librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference (SciPy)*, 2015.

[5] Tim Sainburg. timsainb/noisereduce: v1.0. June 2019.

[6] Tim Sainburg, Marvin Thielk, and Timothy Q Gentner. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS computational biology*, 16(10):e1008228, 2020.

[7] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.