Sentiment Analysis from audio recordings

Abdul Rafay Ahmed Khan
Computer Science
LUMS
Lahore, Pakistan
24100140@lums.edu.pk

Malik M. Moaz
Computer Science
LUMS
Lahore, Pakistan
24100163@lums.edu.pk

Haris Habib
Computer Science
LUMS
Lahore, Pakistan
24020299@lums.edu.pk

***Abstract*—This paper explores the use of machine learning techniques for sentiment analysis on audio recordings using the CREMA-D dataset. The study focuses on analyzing the sentiment of speakers. Various machine learning algorithms, including Support Vector Machines (SVM), Neural Network, K Nearest Neighbours etc, are used to classify the sentiment of the speakers. The study evaluates the performance of the models based on the accuracy of the model. We applied five different models. Results show that the SVM model outperforms all the other models in terms of accuracy and F1 score. With an average accuracy of about 46.5% and average F1 score of about 45%, this model ranks at the top.**

Keywords—sentiment, machine learning, SVM, performance

I. Introduction

Sentiment analysis has become a crucial task in natural language processing (NLP) in recent years. It involves identifying and extracting the sentiment expressed in a piece of text or speech, which can be used for a variety of applications such as understanding customer feedback, monitoring social media sentiment, and predicting stock prices. While sentiment analysis on text has been extensively studied, sentiment analysis on audio recordings is a relatively new area of research. The goal of this study is to explore the use of machine learning techniques for sentiment analysis on audio recordings using the CREMA-D dataset.

The problem of sentiment analysis on audio recordings is important and relevant because it has practical applications in various fields such as market research, psychology, and social media analysis. For instance, analyzing the sentiment of speakers in customer service calls can help identify customer dissatisfaction and improve the quality of service. Similarly, analyzing the sentiment of participants in psychological studies can provide valuable insights into emotional responses and behavior. However, sentiment analysis on audio recordings is a challenging task due to several factors such as background noise, speaker variability, and lack of standardization in speech patterns. Previous research on sentiment analysis has mainly focused on text-based approaches, and relatively few studies have explored sentiment analysis on audio recordings. However, some recent studies have shown promising results in this area. For instance, a study [1] proposed a Speech Emotion Recognition(SER) architecture named DCTFB in which multiple machine learning and deep learning algorithms are used such as SVM, KNN, Logistic Regression, Deep CNN etc which resulted in an accuracy of 86.86% on the SUBESCO dataset. Another study [2] shows the use of Support Vector Machines model in audio recognition using MFCC, PLP, FBANK etc, for feature extraction which results in good accuracy. Shruti et al. [3] proposed a comparative study for sentiment analysis of the Bengali language between ML models and other state of the art models and the study resulted in ML models performing the best. In this study, we aim to build on this previous research and explore the use of different machine learning algorithms for sentiment analysis on audio recordings.

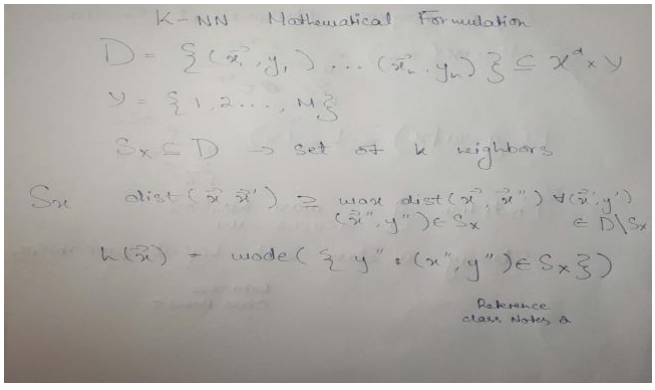
In conclusion, sentiment analysis on audio recordings is an important and challenging problem with many practical applications. While previous research has shown promising results, there is still much to be explored in this area. In this paper, we propose to use machine learning techniques to classify the sentiment of speakers in audio recordings using the CREMA-D dataset. We will evaluate the performance of different machine learning algorithms in this paper.

II. Mathematical Formulation

The given task requires us to classify the emotion expressed by analysing the speech pattern of the speaker. Our goal is to map the features extracted from the sound and map them to label.
For X = {x_1, x_2, ..., x_n} in the feature space we wish to employ an algorithm that maps from X to Y where Y = {y_1 , y_2 , ..., y_n} i.e the label space. For the successful completion of this task however, the following algorithms have been used:
1. K nearest neighbors
2. Logistics Regression
3. Support Vector Machine
4. Naïve Bayes
5. Neural Network (MLP Classifier)

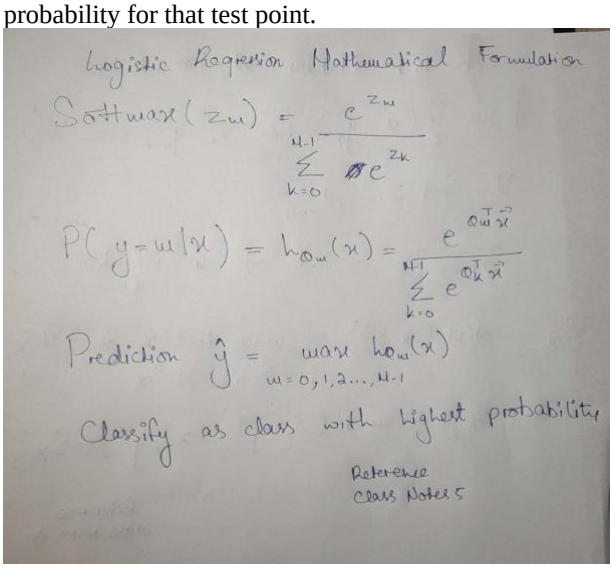
A. K nearest Neighbors:

The K nearest neighbor is a non parametric and instance based classifier. In other words, for it to classify , it need not learn any parameters. Rather it depends on the surrounding neighbors of a test point to classify it . Let us define what neighbors mean. Given a test point its neighbors are all the other training points that lie in the feature space. The classification on the test point is therefore dependent on its K nearest neighbors i.e. those that it is closest to based on a distance metric such as L2 norm.



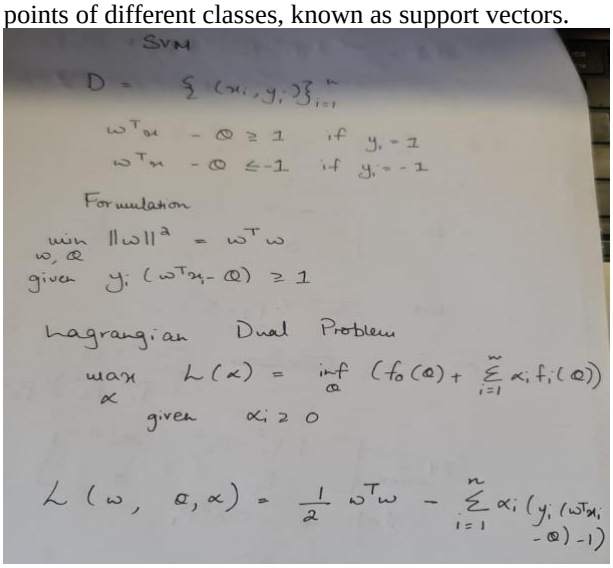
B. Logistic Regression:

Logistic regression relies on first mapping the features from the real line to a probability space i.e [0,1] . It then classifies the point based on whichever class has the highest probability for that test point.



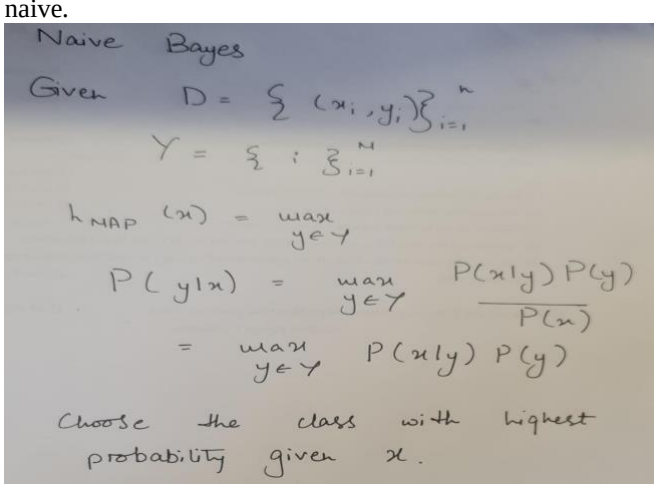
C. Support Vector Machine:

Support Vector Machines (SVMs) are a type of supervised machine learning algorithm used for classification and regression tasks. The goal of an SVM is to find a hyperplane that separates the data points into different classes with the largest possible margin between them. In other words, the algorithm tries to find the best line or plane that separates the data by maximizing the distance between the closest points of different classes, known as support vectors.



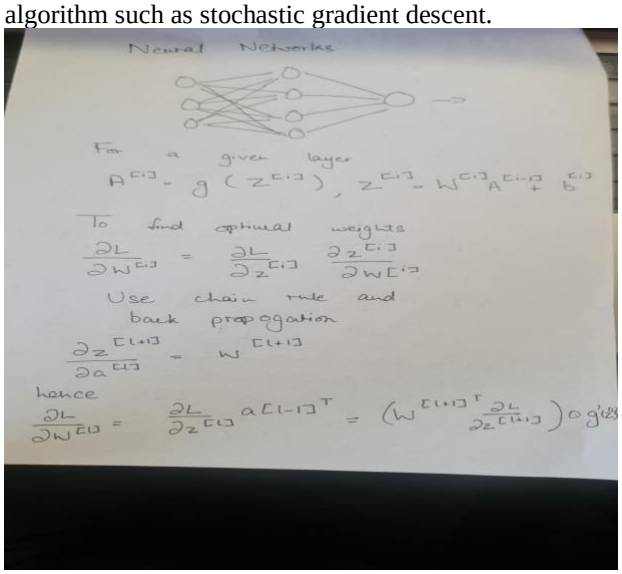
D. Naïve Bayes:

A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the Bayes theorem. Using Bayes theorem, we can find the probability of an event happening, given that another event has occurred. Here, the latter event is the evidence and and the former event is the hypothesis. The assumption made here is that the predictors/features are independent. That is presence of one particular feature does not affect the other. Hence it is called naive.



E. Neural Network:

A neural network is a classifier in which each neuron in the network receives input from other neurons, performs a simple computation on that input, and produces an output. The inputs are multiplied by weights and summed together, and then a non-linear activation function is applied to the result. This output is then sent to the next layer of neurons as input. The weights of the network are learned through a process called backpropagation, which involves adjusting the weights to minimize the difference between the predicted output and the actual output on a training set. This process is typically performed using an optimization algorithm such as stochastic gradient descent.



III. Extraction of Features

Our initial features that we have identified are the following:

1. Mel Frequency Cepstral Coefficients (MFCCs)
2. Spectral Centroids
3. Bispectrum
4. Chromagram
5. Spectrogram

The features we used are the following:

1. Mel Frequency Cepstral Coefficients (MFCCs)
2. Spectral Centroids
3. Chromagram

A. *Mel Frequency Cepstral Coefficients (MFCCs):*

[4] They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal spectrum. This frequency warping can allow for better representation of sound, for example, in audio compression that might potentially reduce the transmission bandwidth and the storage requirements of audio signals.
[5] MFCCs are commonly derived as follows:
Take the Fourier transform of (a windowed excerpt of) a signal.

- Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows or alternatively, cosine overlapping windows.
- Take the logs of the powers at each of the mel frequencies.
- Take the discrete cosine transform of the list of mel log powers, as if it were a signal.
- The MFCCs are the amplitudes of the resulting spectrum.

B. *Spectral Centroids :*

In digital signal processing, spectral centroid is a feature that describes the "center of gravity" of the frequency content of a signal. It is computed by taking the weighted mean of the frequency spectrum of the signal, with each frequency bin weighted by its magnitude. Spectral centroid has been widely used in audio and music analysis for various tasks such as timbre classification, instrument recognition, and genre classification. For example, spectral centroid has been used to distinguish between different musical instruments based on their spectral characteristics [6].

C. *Chromagram:*

A chromagram measures the pitch classes of a sound over time. The pitch classes include A, A#, B, C, C#, D, D#, E, F, F#, G and G#. They can be computed using signal processing techniques such as the short time Fourier transform or the constant Q transform. And then they are mapped to their respective pitch classes.

IV. Feature Engineering

A. *Principal Component Analysis:*

To capture the most discriminating features we have used PCA to reduce features. By virtue of PCA, the projection onto the new feature space has maximum variance.

B. *Variance Threshold(Pseudo):*

To remove the features having zero or near zero variance, variance threshold is used. It essentially measures the variance of all the existing features and eliminates those that do not meet a certain threshold. We have set a threshold of 1.

C. *Pearson Correlation Cofficient:*

To extract features which cause maximum variance in the output, we have calculated the correlation of each feature with the output and include only those which have the highest correlation. We have extracted the top 30 features.

D. *Recursive Feature Elimination:*

Given a certain model, in our case it is Logistic Regression, the algorithm recursively eliminates features based on their model scores in the model.

V. Results

We have summarized the accuracy and f1 scores of all the models after applying all the feature engineering techniques in the tables below. Each row depicts the scores after applying the aforementioned features engineering techniques:

A. *Accuracy Table:*

KNN	Logistic Regression	SVM	Naïve Bayes	Neural Network
41%	42%	45%	36%	38%
37%	43%	45%	33%	38%
42%	45%	48%	37%	41%
41%	41%	45%	39%	39%

B. *F1 Scores Table:*

KNN	Logistic Regression	SVM	Naïve Bayes	Neural Network
41%	41%	44%	34%	38%
36%	42%	43%	31%	38%
41%	43%	47%	34%	41%
40%	40%	45%	38%	39%

These tables show that the SVM model outperforms all the other models in terms of accuracy and f1 scores and is the best classifier.

VI. Conclusion

The aim of this project was to classify emotions based on sound recordings through various Machine learning models. Upon first inspection, we noticed that there was noise in the sound. To remove this noise, we used low pass filteration

and a cutoff frequency of 4000Hz. Upon extracting the clean data, we used various feature extraction techniques and used the models to predict the emotions. The results show that the features that we used to run the models for predictions do not explain the variation in the output completely even though we tried various techniques, but the scores could not increase much. For future research, there is still potential features to be identified that could explain the output much better and result in better scores.

REFERENCES

[1] S. Sultana, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, "Sust bangla emotional speech corpus (subesco): An audio-only emotional speech corpus for bangla," Plos one, vol. 16, no. 4, p. e0250173, 2021.

[2] I Nattapong Kurpukdee , Sawit Kasuriya , Vataya Chunwijitra ,Chai Wutiwiwatchai and Poonlap Lamsrichan ," A Study of Support Vector Machines for Emotional Speech Recognition", 978-1- 5090-4809-0/17/$31.00 ©2017 IEEE

[3] A. C. Shruti, R. H. Rifat, M. Kamal and M. G. R. Alam, "A Comparative Study on Bengali Speech Sentiment Analysis Based on Audio Data," 2023 IEEE International Conference on Big Data and Smart Computing (BigComp), Jeju, Korea, Republic of, 2023, pp. 219-226, doi: 10.1109/BigComp57234.2023.00043.

[4] Min Xu; et al. (2004). "HMM-based audio keyword generation" (PDF). In Kiyoharu Aizawa; Yuichi Nakamura; Shin'ichi Satoh (eds.). Advances in Multimedia Information Processing – PCM 2004: 5th Pacific Rim Conference on Multimedia. Springer. ISBN 978-3-540-23985-7.

[5] Sahidullah, Md.; Saha, Goutam (May 2012). "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition". Speech Communication. 54 (4): 543–565.

[6] McAdams, S., & Winsberg, S. (1993). Natural constraints and musical shape. Computer Music Journal, 17(3), 24-33.