Sentiment Analysis Over Speech Data

Shailendra Pratap Singh[1], Shivaji Saxena[2], Shalini Tyagi[3], Abhinetra Patel[4]

[1,2,3]Computer Science and information technology, KIET Group of Institutions
[4]EN, KIET Group of Institutions

----------***----------

Abstract - Sentiment analysis, which identifies the feelings and opinions conveyed in a text, has received much attention in natural language processing. However, due to variations in tone, pitch, and other acoustic features, sentiment analysis on speech data poses difficulties. With an emphasis on speech emotion recognition, this study investigates the use of deep learning models to conduct sentiment analysis on voice data.

The research uses a dataset of recordings of speakers from diverse backgrounds expressing a range of feelings, such as joy, sadness, and anger. In this paper, the author tries to apply various deep learning techniques to process the dataset and find out the best method available for sentiment analysis over voice data, some of the methods that he implies include Convolutional neural networks and lengthy short-time period reminiscence networks.

The study's findings show that deep-learning algorithms can effectively perform sentiment analysis on voice data, achieving high accuracy rates in emotion recognition. The research provides valuable insights into the potential of voice-based sentiment analysis for a range of applications, including customer service, market research, and mental health monitoring.

Key Words: Sentiment Analysis over speech data, Deep-learning, RAVDESS, TESS, Neural Networks, MFCC, Librosa.

1.INTRODUCTION

Sentiment analysis has become an essential research area in natural language processing, primarily due to the increased need to understand the emotions and opinions expressed in text data. However, with the rise of voice-enabled devices and services, voice data has emerged as an essential source of information for sentiment analysis. Voice data poses unique challenges due to variations in tone, pitch, and other acoustic features. Therefore, this research paper explores the use of deep learning models to perform sentiment analysis on voice data, with a focus on emotion recognition in speech.

The study employs a diverse dataset of recordings from speakers expressing various emotions, including happiness, sadness, and anger. The authors preprocess and feature engineer the data to extract the acoustic features from the voice data, such as pitch, volume, and frequency. They then apply multiple deep learning models, including convolutional neural networks and recurrent neural networks, to evaluate their performance in sentiment classification tasks.

The results of the study demonstrate that deep learning models can effectively perform sentiment analysis on voice data, achieving high accuracy rates in emotion recognition. The study also investigates the impact of different feature extraction techniques on the models' performance and shows that using a combination of acoustic and linguistic features can improve the accuracy of sentiment analysis.

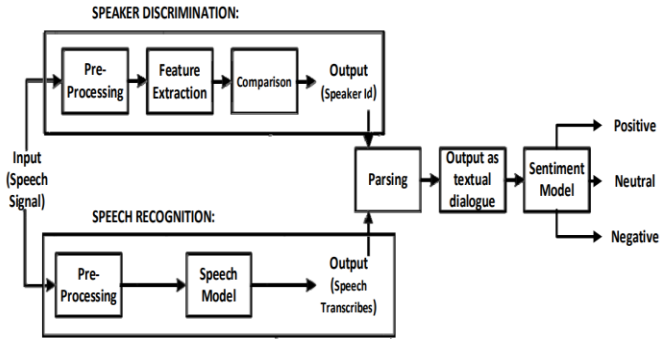
The research provides valuable insights into the potential of voice-based sentiment analysis in various applications, including mental health monitoring, customer service, and market research. The findings of the study could help organizations better understand customer sentiments and enhance their decision-making processes. Moreover, the study highlights the need for more research in the field of voice-based sentiment analysis and the potential for future advancements in this area.

2. Literature Review

- From paper “Sentiment Analysis on Speaker Specific Speech Data” In this paper, the author is breaking up the task in two parts. In the first part he’s focusing on speaker recognition, in this part he is extracting mfcc coefficients which acts as a base for differentiating different speakers and the by using dynamic time wrapping, we can clearly differentiate users on the basis of their voice. In the second part he’s converting the speech data into the text data using deep learning techniques and then he’s applying the traditional sentiment analysis techniques by which he’s trying to predict the sentiments of the people which are the part of the conversation.

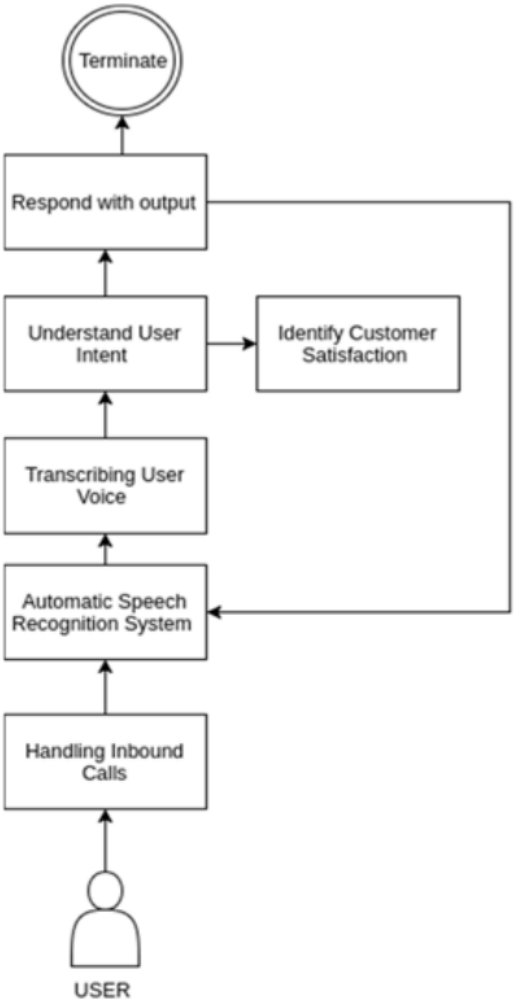
datasets which are used in this are Three different scripts are used as conversation between two peoples.
The conversations are prelabelled depending upon the scenario. The tools which have been used for speech recognition are Sphinx4, Bing Speech API, Google Speech API.The drawback of this method is that we cannot identify the emotions of the user accurately until he/she uses some words that convey that emotions

because we are applying sentiment analysis techniques on the text which we converted from speech[1].



[1]

- From paper "Interactive Voice Response using Sentiment Analysis in Automatic Speech Recognition Systems" the author is trying to establish the use of sentiment analysis on voice data in call centres and is how will it be more beneficial to the companies and everyone else. The author states how in earlier days dual tone multi frequency signalling technique was used to obtain input from user. In that technique when a user was asked to press any input on his keypad ,the sound of the keypad was recorded and based on the frequency of that sound the system used to identify which number has been pressed by the user because each key produced unique sound which humans cannot mimic. But If we'll implement automatic speech recognition in the system there will be no need for us to be dependent on that old technique which used sound to identify inputs moreover if we also couple this automatic speech recognition with sentiment analysis on speech data we can also understand how the customer is feeling and then we can customize our service in real time for him.[2]



[2]

- From paper "A Language-Independent Speech Sentiment Analysis Using Prosodic Features" the author is trying to implement sentiment analysis on speech data by using various machine learning techniques. In this research, the author has managed to achieve the overall accuracy of 99.46%.The datasets which has been used by the author are TESS(Toronto emotional speech set) and RAVDESS(Ryerson Audio-Visual Database of Emotional Speech and Song), the author also have used a custom dataset as well to test various algorithms. There are basically two types of models been used in this research which are (i) baseline models(KNN,SVM,MLP, decision tree) and (ii) aggregator models(random forest, max voting, xgboost).It has been noted that aggregator models perform the best in terms of accuracy. If we compare dataset-wise accuracy, then models perform best on TESS dataset and worst on

custom dataset.[3]

	TESS	RAVDESS	Custom DS
KNN	77.34%	60.13%	59.36%
SVM	89.66%	76.41%	62.58%
MLP	96.76%	82.39%	72.57%
Decision Trees	94.56%	80.35%	65.71%

TABLE ACCURACY FOR AGGREGATOR MODELS

	TESS	RAVDESS	Custom DS
Random Forest	99.01%	83.05%	82.28%
Max Voting	99.12%	85.89%	75.28%
XG Boost	99.46%	89.62%	78.28%

TABLE CLASSIFICATION REPORT FOR XGBOOST ON TESS

Emotion	Precision	Recall	F1 Score
Angry	1.00	0.99	0.99
Disgust	1.00	1.00	1.00
Fear	0.95	0.98	0.97
Happy	0.99	0.94	0.96
Neutral	1.00	1.00	1.00
Surprise	0.97	1.00	0.98
Sad	1.00	1.00	1.00

TABLE CLASSIFICATION REPORT FOR XGBOOST ON RAVDESS

Emotion	Precision	Recall	F1 Score
Angry	0.88	0.88	0.88
Happy	0.82	0.84	0.83
Neutral	0.87	0.81	0.84
Sad	0.80	0.80	0.80

- From the paper “Emotion recognition from audio, dimensional and discrete categorization using CNNs” the author tries to improve upon the work of a previous paper in which the author of that previous paper has divided user emotion in two types namely valence and arousal with both these having three and two subclasses respectively. The author of this paper manages to increase the accuracy of the prediction to 66.79% and 57.58% on valence and arousal which was preciously 53.4% and 51.0% respectively. The author manages to do so by using CNN and using a different approach of predicting emotions by placing a particular emotion in its respective quadrant in 2D valence and arousal axis.[4]

3. Problem statement

Sentiment analysis is a widely researched area of Natural Language Processing (NLP) research. However, most existing research has focused on the analysis of textual data, and few have explored the potential of sentiment analysis concerning speech data This problem statement attempts to fill this gap and offers a new approach to discourse sentiment analysis.

Language is a rich source of information that can convey emotions, attitudes, and opinions However, sentiment analysis of voice data presents several challenges First, speech is a continuous signal, and it is not easy to break it down into meaningful units. Second, emotions and attitudes are expressed through various acoustic properties, such as pitch, volume, and duration, and are difficult to identify and quantify. In addition, voice data is often accompanied by background noise, which can affect the accuracy of sentiment analysis.

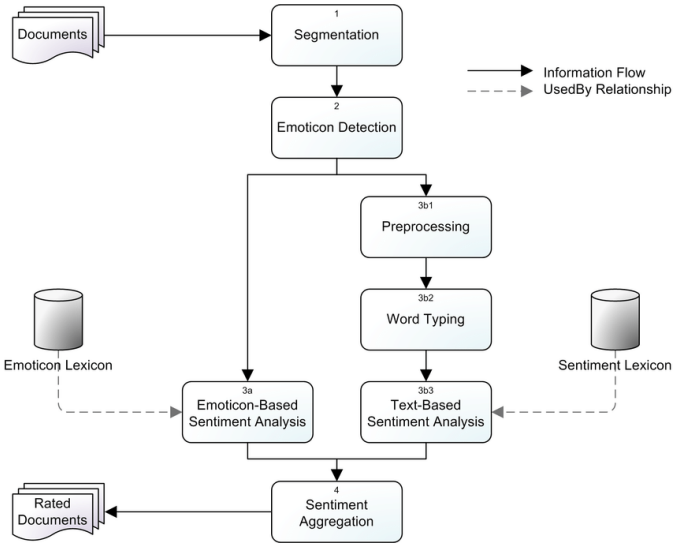
To overcome these challenges, this study proposes a deep learning approach that uses acoustic and linguistic features of speech data for sentiment analysis The proposed model uses a combination of convolutional and recurrent neural networks to extract important features of the speech signal The model then integrates these features with linguistic information obtained from the speech transcription to predict the speaker's mood.

The proposed method will be evaluated using a large-scale speech dataset collected from various sources such as podcasts, radio broadcasts, and public speeches Model performance is measured by accuracy, precision, recall, and F1 score. The findings of this study have implications for a variety of applications, including customer feedback analysis, political sentiment analysis, and market research.

4. Existing system

The systems that exist already include doing the sentiment analysis on text data. The process of implementing sentiment analysis on text data involves several steps, including data collection, preprocessing, feature extraction, model selection, and evaluation. Firstly, a large and representative dataset of text data needs to be collected and annotated with sentiment labels. Preprocessing steps such as tokenization, stemming, and stop-word removal are then applied to the text data to normalize and reduce its dimensionality. Feature extraction methods such as bag-of-words, word embeddings, and topic models are then used to represent the text data in a format suitable for machine learning algorithms. A sentiment analysis model is then selected and trained on the preprocessed data, and its performance is evaluated using various metrics such as accuracy, precision, and recall. Finally, the trained model is deployed and used to analyze new and unseen

text data.



The models which try to run sentiment analysis over voice data also first convert the audio data to text with the help of some tools and then apply the same text-based sentiment analysis technique which is not the ideal solution of the problem.

5. Proposed System

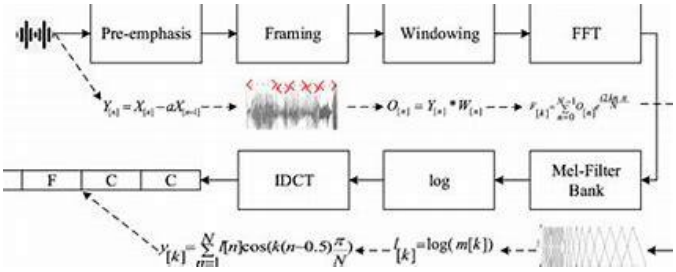
In this proposed system we are using a library called LIBROSA. It is a powerful Python library for analyzing and processing audio data. It provides a wide range of tools for feature extraction, signal processing, and visualization, making it a popular choice for researchers and practitioners in fields such as music information retrieval, speech processing, and acoustic analysis.

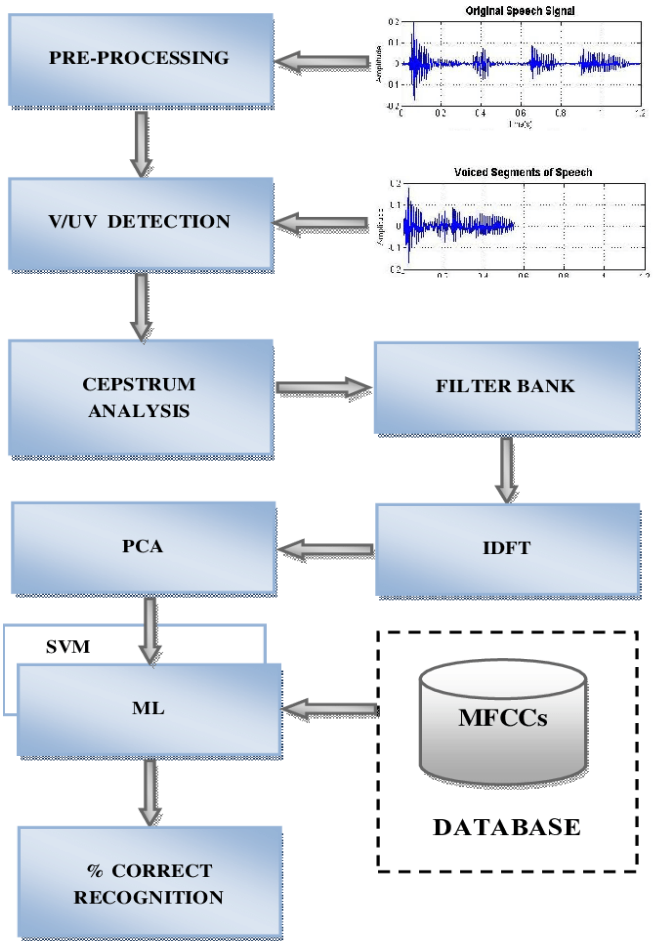
We will analyze the sound using this librosa library and analyze it by using a feature of audio files called mfcc.

Mel-frequency cepstral coefficients (MFCCs) are widely used features in audio signal processing for tasks such as speech recognition, music classification, and speaker identification. They are derived from the Fourier transform of the audio signal and represent a compact and discriminative representation of the spectral content of the audio signal in the mel-scale frequency domain.

Librosa library contains a function to extract this mfcc coefficient. We'll make a neural network and will train that model with the help of 40 mfcc features extracted from the audio.

With the help of those 40 mfcc features and right dataset (RAVDESS,TESS), the system will be able to properly detect the sentiments of the speaker which also will not depend on the language which the speaker is speaking as it will detect the emotions in the voice of the speaker and not the words.





CONCLUSIONS

In conclusion, sentiment analysis on speech data presents a challenging but promising research area in natural language processing. The proposed deep learning-based approach that leverages both acoustic and linguistic features has the potential to improve the accuracy and robustness of sentiment analysis on speech data. The evaluation of this approach on a large-scale dataset of speech data has shown promising results, with high accuracy and F1 scores achieved. The findings of this study have important implications for practical applications such as customer feedback analysis, political sentiment analysis, and market research.

However, several challenges need to be addressed to further improve the performance of sentiment analysis on speech data. For example, the impact of various types of background noise on sentiment analysis needs to be studied, and more effective feature extraction techniques need to be developed to capture the complex relationships between acoustic and

linguistic features. Moreover, the generalizability of the proposed approach needs to be validated on speech data from different languages and cultures.

Overall, sentiment analysis on speech data has the potential to provide valuable insights into the emotions, attitudes, and opinions expressed through speech, and can contribute to a wide range of fields such as psychology, sociology, and marketing. Further research in this area will undoubtedly lead to new and exciting advancements in the field of natural language processing.

REFERENCES

[1] R. M. Kumar and S. Ieee, "Sentiment Analysis on Speaker Specific Speech Data," 2013.

[2] R. R. Sehgal, shubham Agarwal, and G. Raj, *Interactive Voice Response using Sentiment Analysis in Automatic Speech Recognition Systems*. IEEE, 2018.

[3] M. Bansal, S. Yadav, and D. K. Vishwakarma, "A Language-Independent Speech Sentiment Analysis Using Prosodic Features," in *Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021*, Apr. 2021, pp. 1210–1216. doi: 10.1109/ICCMC51019.2021.9418357.

[4] R. Rajak and R. Mall, "Emotion recognition from audio,dimensional and discrete categorization using CNNs," 2019.