

Speech to Text Conversion and Sentiment Analysis on Speaker Specific Data

Submitted by:
Sneha Dutta

Completion Date:

September 14,2024

Personal Project



Sardar Vallabhbhai National Institute of Technology, Surat,
Gujarat-395007

ABSTRACT

In this project, development of a pipeline for emotion detection from speech data, starting with speech-to-text conversion using the Google Speech Recognition API, followed by sentiment and emotion classification is being done. A pre-trained DistilRoBERTa model from Hugging Face was used to classify emotions like joy, sadness, and anger from the transcribed text.

The model's performance is being evaluated using accuracy, precision, recall, and F1-score, achieving 79% accuracy on a dataset of 40 data points. Performance varied across emotion classes due to dataset size and class imbalance. Future improvements will focus on expanding the dataset and balancing emotion categories to enhance model generalization.

Contents

Index	ii
1 Introduction	1
2 Methodology	2
3 Results and Discussions	4
3.1 Speech Recognition	4
3.2 Polarity and Subjectivity	5
3.3 Final Result	6
4 Conclusion	8

Chapter 1

Introduction

Speech recognition is a technology that enables machines to interpret and convert human speech into text. It plays a crucial role in various applications, such as virtual assistants, voice-controlled devices, transcription services, and natural language processing tasks. One popular API for speech-to-text conversion is the Google Speech Recognition API, which can recognize speech from various languages and accents.

In this project, the goal is to convert speech into text using the Google Speech Recognition API. The collected speech data from multiple speakers will be processed through this API to generate textual output. After the speech is converted into text, the next step is to perform sentiment analysis on the text.

Sentiment analysis is a task in Natural Language Processing (NLP) that involves determining the emotional tone behind a body of text. Common models for sentiment analysis include Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNN), which are widely used for handling sequential data.

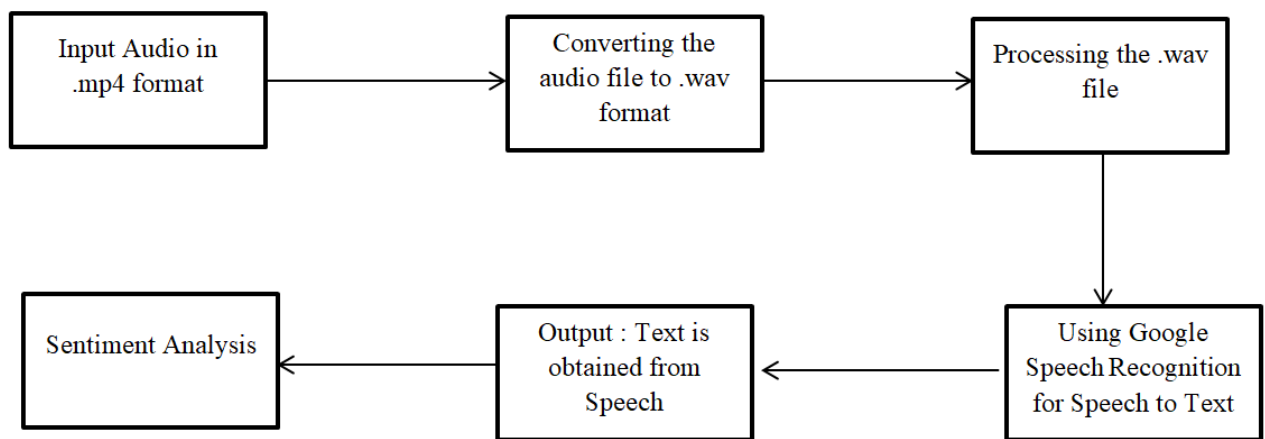
However, in this project, we are dealing with a small dataset consisting of only 40 samples of speech data collected from various speakers. Due to this limitation, using complex models like LSTM or RNN may lead to overfitting, where the model performs well on training data but fails to generalize on new data.

To address this, instead of using LSTM, the project will utilize TextBlob, a Python library designed for text processing, including sentiment analysis. TextBlob is pre-trained and simple to use, making it suitable for small datasets. It can efficiently handle the limited data and provide accurate sentiment predictions without the risk of overfitting. TextBlob is an ideal choice for this project due to its ease of implementation and effectiveness in small-scale sentiment analysis tasks.

Chapter 2

Methodology

This section describes the project's structured methodology. The two primary components of the process are Speech-to-Text Sentiment Analysis and Conversion. The project workflow is visually represented by the block diagrams, and each stage is described in more detail below.



Block diagram of the system

1.Speech-to-Text Conversion

Speech Data Collection:

The first step involves collecting audio data from multiple speakers. The dataset consists of 40 audio recordings in .wav format, each representing a sentence or a short paragraph spoken by individuals. These audio recordings are gathered from different speakers with varying accents and tones to ensure diversity in the dataset.

Google Speech Recognition API: The collected audio files are processed through the Google Speech Recognition API to convert them into text. The API listens to the speech and outputs a corresponding textual representation.

Preprocessing of Audio:

Before feeding the audio files to the API, the recordings are checked for quality (e.g., removing noise or silence), and the appropriate file format is ensured (.wav).The converted text output from each audio file is stored for further processing.

2.Sentiment Analysis

TextBlob Library:

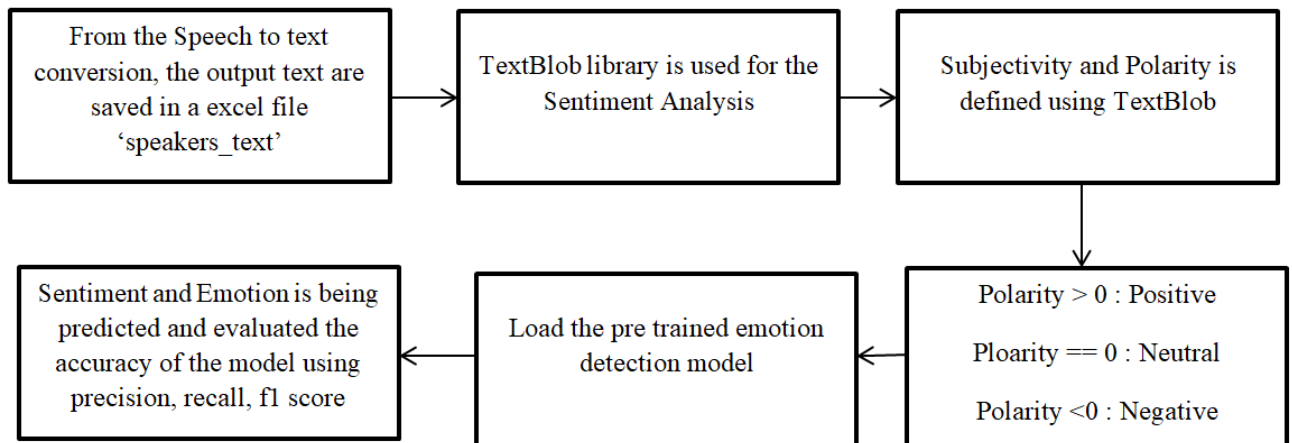
After the speech has been converted into text, TextBlob, a Python library for text processing, is used to analyze the sentiment of each piece of text. TextBlob is pre-trained and capable of determining the polarity (positive, negative, neutral) and subjectivity of the text.

Justification for Using TextBlob:

Given that the dataset is relatively small, with only 40 samples, using complex machine learning models such as LSTM or RNN would lead to overfitting. As a result, TextBlob is selected for its efficiency with small datasets, simplicity, and reliability in performing sentiment analysis.

Sentiment Output:

TextBlob processes each sentence and returns a sentiment score, which is categorized into positive, negative, or neutral. The sentiment scores are stored and compared across different speakers to analyze the overall emotional tone of the dataset.



Block diagram of the Sentiment Analysis

Chapter 3

Results and Discussions

3.1 Speech Recognition

Speech recognition, or speech-to-text, is the ability of a machine or program to identify words spoken aloud and convert them into readable text. Speech recognition uses a broad array of research in computer science, linguistics and computer engineering. Many modern devices and text-focused programs have speech recognition functions in them to allow for easier or hands-free use of a device.

Different speakers' input is gathered as indicated below. The audio sample, which was recorded in a quiet, controlled setting, displays the pitch and loudness of their voice. The samples were further converted into text and the length of time it took the speaker to pronounce each sentence are displayed in the following table.

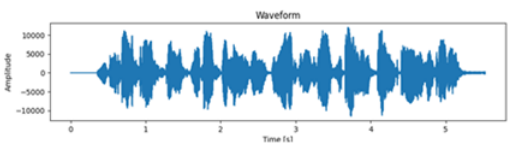
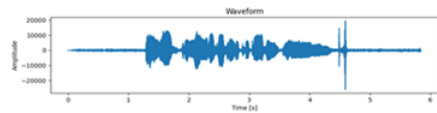
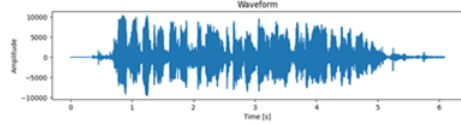
Speaker	Audio : Input	Text Sentence	Time(seconds)
Speaker1		I want to buy track pant but after visiting two or three shops I didn't find any which I like	5.52
Spekaer2		I have too many lug gages to carry on my own	5.80
Spekaer3		oh I totally forgot that I have to complete the assignment and tomorrow is the submission	6.08

Figure 3.1: Input audio to text conversion

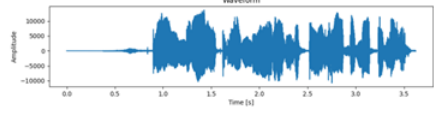
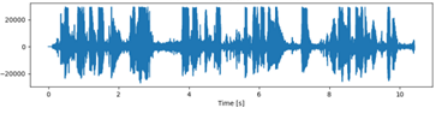
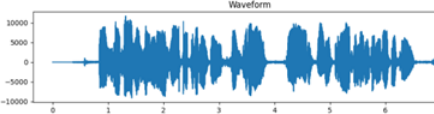
Speaker	Audio : Input	Text Sentence	Time(seconds)
Spekaer4		trying out a new recipe for dinner	3.62
Spekaer5		laughter is the key to joy I got this thought after attending a stand up comedy show today	10.32
Spekaer6		everyday is a new opportunity to grow learn and create something beautiful	7.10

Figure 3.1: Input audio to text conversion

3.2 Polarity and Subjectivity

In sentiment analysis, polarity and subjectivity are two key components used to interpret the emotional tone of a piece of text. They help in categorizing and understanding opinions expressed in the text.

Polarity: Polarity refers to the sentiment orientation or emotional direction of a text. It is the measure of how positive, negative, or neutral a given statement is.

Favorable or upbeat sentiments are indicated by positive polarity, as in the case of remarks like "I love this product." On the other hand, negative polarity conveys an unfavorable or pessimistic feeling, such as "This service is terrible." Conversely, neutral polarity conveys an impartial or well-rounded viewpoint devoid of overtly positive or negative feelings, as in the statement "The product is red."

In most cases, polarity values are shown on a scale. The continuous scale, with values ranging from -1 (very negative sentiment) to +1 (extremely positive sentiment), is one popular approach. As an alternative, polarity can be divided into distinct classes—usually three or more—that classify emotions into categories that are positive, negative, and neutral.

Subjectivity: Subjectivity refers to the degree of personal opinion or emotional expression in a text. It measures how much of the text reflects subjective viewpoints versus objective facts.

Subjective text: Contains personal opinions, feelings, or beliefs (e.g., "I think this movie is amazing"). These statements often carry sentiment and personal judgment.

Objective text: States facts or information without personal feelings (e.g., "The product was released in 2020"). These statements typically lack any sentiment expression.

Subjectivity values are often measured on a scale:

0 indicates completely objective, fact-based text.

1 indicates highly subjective text filled with opinions and emotions.

Speaker	Input:Text	Polarity	Subjectivity
Speaker1	I want to buy track pant but after visiting two or three shops I didn't find any which I like	0.000000	0.000000
Speaker2	the state of corruption in our society is utterly disgusting	-1.000000	1.000000
Speaker3	oh I totally forgot that I have to complete the assignment and tomorrow is the submission	0.050000	0.575000
Speaker4	trying out a new recipe for dinner	0.136364	0.454545
Speaker5	laughter is the key to joy I got this thought after attending a stand up comedy show today	0.400000	0.600000
Speaker6	everyday is a new opportunity to grow learn and create something beautiful	0.262121 3	0.684848

Table 3.1: Polarity and subjectivity

3.3 Final Result

The table below depicts the overall sentiment and emotion of the text of the input speech. The sentiments depicted are positive, negative and neutral in these cases.

Performance Analysis : Precision, recall and f1 score is being used to evaluate the performance of the sentiment analysis model.

Precision: Precision is the proportion of true positive predictions out of all positive predictions made by the model. It tells you how many of the predicted positive instances are actually correct.

Recall: Recall (also known as sensitivity or true positive rate) is the proportion of true positives out of all the actual positives. It shows how well the model identifies true positives.

f1-score: The F1-score is a performance metric used to evaluate the accuracy of a classification model. It is the harmonic mean of precision and recall, providing a single score that balances the two.

Speaker	Text	Sentiment	Emotions
Speaker1	I want to buy track pant but after visiting two or three shops I didn't find any which I like	Neutral	sadness
Speaker2	the state of corruption in our society is utterly disgusting	Negative	disgust
Speaker3	oh I totally forgot that I have to complete the assignment and tomorrow is the submission	Positive	surprise
Speaker4	trying out a new recipe for dinner	Positive	Neutral
Speaker5	laughter is the key to joy I got this thought after attending a stand up comedy show today	Positive	joy
Speaker6	everyday is a new opportunity to grow learn and create something beautiful	Positive	joy
Speaker7	feeling bitter about the unfairness in the job place	Negative	anger
Speaker8	the feeling of the unknown is keeping me up at night	Negative	fear

Table 3.2: Overall Sentiment of the speaker

Accuracy: 79.49%				
	precision	recall	f1-score	support
anger	1.00	1.00	1.00	1
disgust	1.00	1.00	1.00	2
fear	1.00	0.67	0.80	3
happiness	0.00	0.00	0.00	2
joy	0.82	0.82	0.82	17
neutral	0.80	0.80	0.80	5
sad	0.00	0.00	0.00	1
sadness	0.78	1.00	0.88	7
surprise	0.33	1.00	0.50	1
accuracy			0.79	39
macro avg	0.64	0.70	0.64	39
weighted avg	0.76	0.79	0.77	39

Depicts the accuracy along with other evaluating parameters

Chapter 4

Conclusion

This project was developed from scratch, beginning with the collection of speech data from various speakers. The audio recordings were processed and transcribed into text using Google's Speech Recognition system. For sentiment analysis, the pre-trained Python library, TextBlob, was utilized. The analysis evaluated the subjectivity, polarity, and emotions present in the text data generated from the speech inputs.

To assess the performance of the model, precision, recall, and F1-score were used as evaluation metrics. The model achieved an accuracy of 79.49