Audio Emotion Analysis

Anudeep Nayak
Kalp Mepani

Introduction:

Emotions are an important part of communication. Words can have different meanings depending on emotions. This project focuses on developing a system that converts audio to text and subsequently analyzes the emotional undertones of the spoken word. Utilizing the capabilities of the Speech-to-Text (STT) models and Natural Language Processing (NLP), this project aims to offer insights into the emotional aspect of spoken communication.

Background:

Nobert Wiley in his paper “Emotion and Film Theory” claims that Everyday emotion is loose in frame or context but rather controlled and regulated in content. Movie emotion, in contrast, is tightly framed and boundaried but permissive and uncontrolled in content. Movie emotion is therefore quite safe and inconsequential but can still be unusually satisfying and pleasurable.
This project chooses a movie to do audio analysis to understand the boundaried emotions as mentioned by Nobert Wiley.

Objectives:

This project aims to answer several questions. Such as

- Do sentiments (opinion or view) differ from emotions?
  Hypothesis: It should differ as some emotions can be positive but have a negative sentiment.
- Are both these the same or different?
  Hypothesis: Both these emotions and sentiments are different
- How does emotion evolve throughout the movie?
  Hypothesis: Since the movie is about Anne Frank, it should have a lot of sad emotions
- Does the length of the sentence have anything to do with the emotions and sentiments?
  Hypothesis: The length of sentences should be independent of emotions or sentiments, as both of these can be expressed in a single word.
- Do sentences combined have different emotions and sentiments compared to a single sentence?
  Hypothesis: Sentences combined have different emotions and sentiments.

Data used:

Two datasets have been used to complete the analysis. The first one is an animated movie called “ANNE FRANK’s DIARY- Animated feature film”, this movie is taken from YouTube, and is

then converted to an MP3 file. The Whisper AI transcribes the MP3 file to text, this text is broken down into sentences and converted to a data frame, which is then used to analyze the emotions inscribed inside the texts.

The second dataset “Emotions dataset for NLP” by Praveen is taken from Kaggle to build a model that recognizes emotions embedded within the texts.

Interesting fact about emotions, there are 7 basic emotions anger, sadness, fear, surprise, disgust, contempt, and happiness. This dataset uses 5 of these emotions.

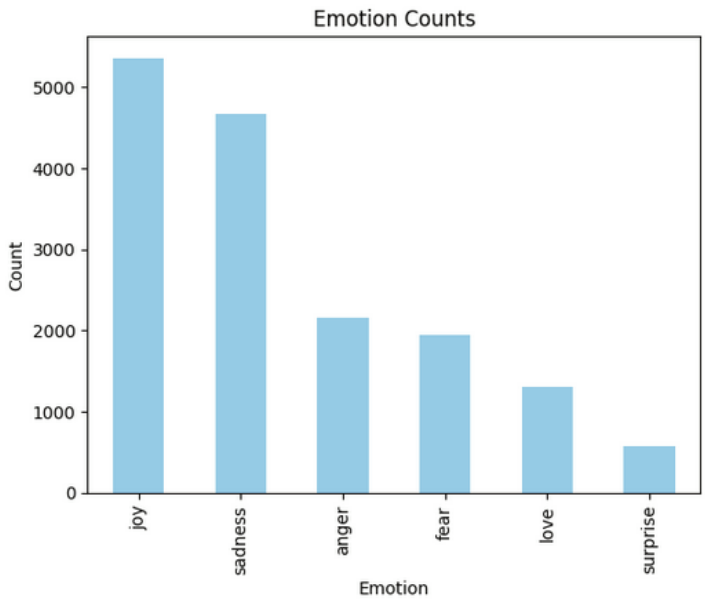

Fig 4.1

The dataset looks like the following, with “Text” as a description of the text and “Emotions” as the true emotions of the given texts.

	Text	Emotions
0	i can go from feeling so hopeless to so damned...	sadness
1	im grabbing a minute to post i feel greedy wrong	anger
2	i am ever feeling nostalgic about the fireplac...	love
3	i am feeling grouchy	anger
4	ive been feeling a little burdened lately wasn...	sadness

Fig 4.2

Methods and Models:

This project uses multiple models and methods to answer the objectives. First, the audio file is fed to the Whisper AI model to get the transcribed texts, this text is broken down into sentences and is stored as a data frame. Parallelly, the “Emotions dataset for NLP” is taken to train a neural network model, which classifies the input into emotions.
The data frame is fed to the neural network to get the emotion for each sentence. This data frame is again fed into TextBlob, a library in Python to get the sentiment of each sentence. And then, we do the correlation analysis to get some valuable insights and conclusions to our questions

Whisper AI:

Whisper is an automatic speech recognition (ASR) system trained on 680,000 hours of multilingual and multitask supervised data collected from the web. This AI is used to extract text from the movie in this project.

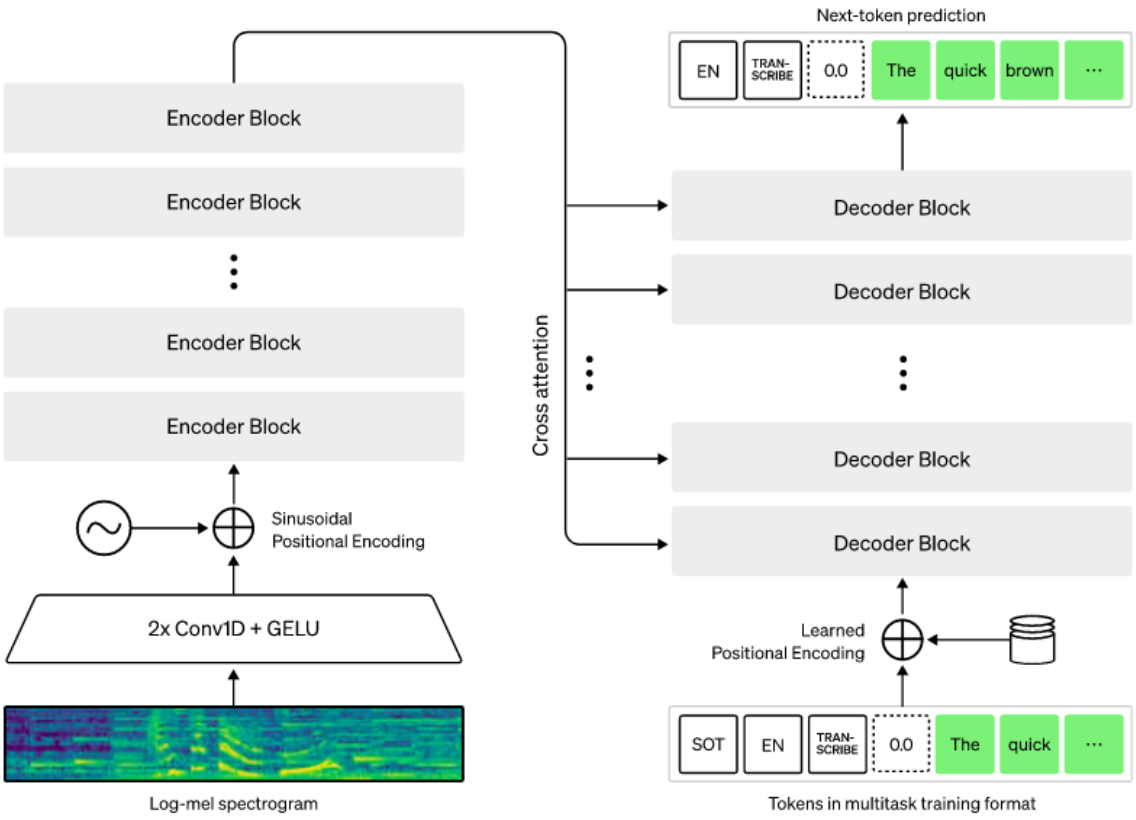

Fig 5.1

The Whisper architecture is a simple end-to-end approach, implemented as an encoder-decoder Transformer. Input audio is split into 30-second chunks, converted into a

log-Mel spectrogram, and then passed into an encoder. A decoder is trained to predict the corresponding text caption, intermixed with special tokens that direct the single model to perform tasks such as language identification, phrase-level timestamps, multilingual speech transcription, and to-English speech translation. This explanation and images are taken from the OpenAI website.

An interesting fact about Whisper AI: It can transcribe at least 57 languages besides English.

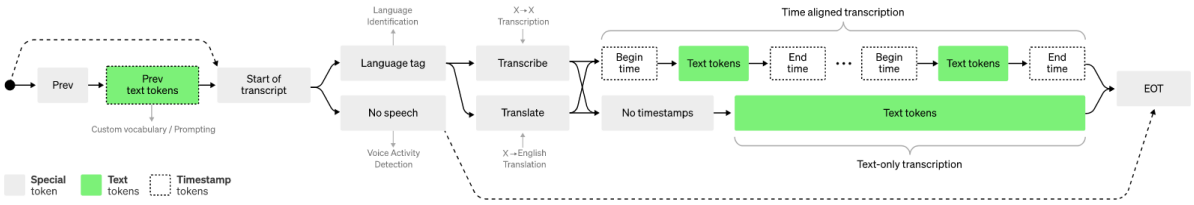

Fig 5.2

Neural Network:

“Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns. They interpret sensory data through a kind of machine perception, labeling or clustering raw input. The patterns they recognize are numerical, contained in vectors, into which all real-world data, be it images, sound, text or time series, must be translated.” This quote is an explanation of the Neural network from the Path mind.
This project uses this concept to get the classification of texts into emotions.

Classification is a supervised machine learning method, which classifies the data into pre-defined categories. This project uses the Neural network to achieve the classification.

The neural network is modeled loosely after the human brain and uses nodes like the human brain to activate a signal. Here is a visual representation of a node,

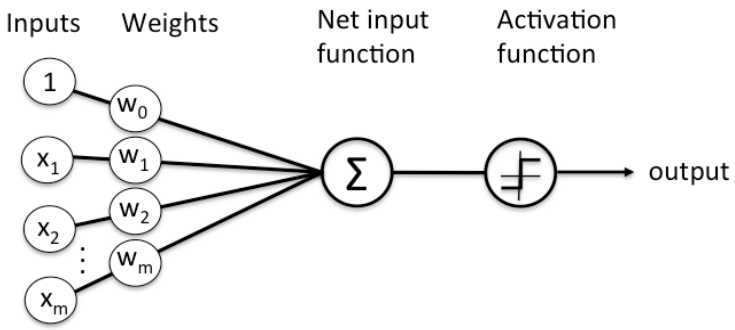

Fig 5.3

A row of such nodes make a layer, many layers make a network. Below is the common representation of a network in a typical Neural network.

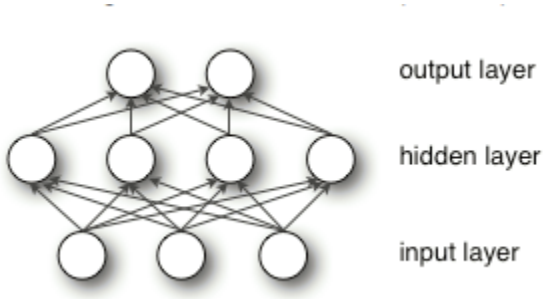

Fig 5.4

Data is passed through the input layer. Each connection between nodes in consecutive layers has a weight associated with it. These weights determine the strength of the connection. After calculating the weighted sum of inputs for each node in the hidden layers and output layer, an activation function is applied to introduce non-linearity. Hidden layers in the network learn more complex patterns of the data, and then the output is predicted from the output layer, which is held against the true value to measure accuracy using the loss function. Backpropagation is performed to adjust weight and train the network again to reduce the loss. When loss is sufficiently reduced, the network is ready for the inputs for prediction.

An interesting fact about the neural network: Its invention can be traced back to 1943, when Warren McCulloch, a neurophysiologist, and Walter Pitts, a logician, laid down the foundational work in their paper titled "A Logical Calculus of Ideas Immanent in Nervous Activity."

TextBlob:

TextBlob is a Python text processing package that offers a streamlined user interface for typical natural language processing (NLP) activities. TextBlob is an easy-to-use package that makes Python jobs related to natural language processing more manageable for users of different skill levels. Common text processing tasks are made easier by it, such as sentiment analysis, translation, noun phrase extraction, part-of-speech tagging, classification, and more. TextBlob is based on the Natural Language Toolkit (NLTK) and builds upon its features by providing an easier-to-use API. Sentiment analysis, which lets users determine a text's sentiment polarity (positive, negative, or neutral), is one of its standout capabilities.

Fig 5.5

So, we gave the input that we got from the Whisper AI model into the TextBlob for getting the sentiments - positive, negative, or neutral and the neural network gave us the sentiments - joy, sadness, anger, fear, love, and surprise.

Correlation Analysis:

A statistical technique for determining the direction and degree of a linear relationship between two quantitative variables is correlation analysis. Put differently, it aids in evaluating the relationship between changes in one variable and changes in another. A correlation analysis yields a correlation coefficient, which expresses how much the variables are associated. It is used to explore and analyze patterns in data by measuring and interpreting the degree of relationship between two quantitative variables. The correlation between 2 variables can be either positive, or negative or they both can be independent.

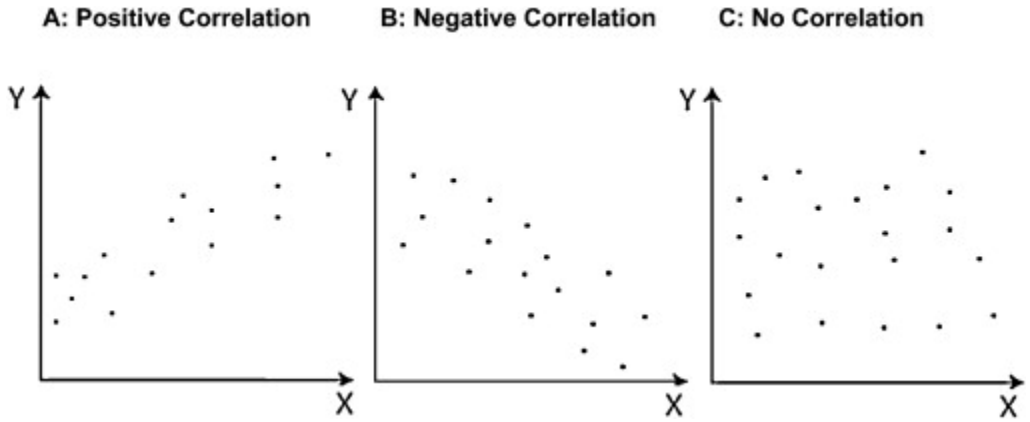

Fig 5.6

Python Libraries and Modules:

- Numpy: An efficient Python numerical library that supports big, multi-dimensional matrices and arrays, as well as mathematical operations on them.

- Keras:
  - Tokenizer: Text is tokenized into words or subwords for tasks involving natural language processing.
  - Sequential: Enables the neural network layers to be stacked linearly.
  - Embedding: Provides dense vectors for word embeddings by converting positive integers.
  - Flatten: To connect convolutional layers to tightly connected layers, flattening the input data is useful.
  - Dense: Uses neural networks with fully connected layers.

- Pandas: A data manipulation package that provides effective data analysis and manipulation tools such as DataFrame data structures.

- Nltk:
  - Sent_tokenize: Splits the text into sentences, which is an important step in natural language processing and text analysis

- Tensorflow:
  - Pad_sequences: This to a specific length, is mostly used in sequence data preprocessing for neural networks

- Sklearn:
  - Labelencoder: This converts the categorical labels into numbers for machine learning models.
  - Train_test_split: This splits the dataset into training and testing sets for the evaluation of the model

- Matplotlib: This was mainly used by us in creating static, animated, and some interactive visualizations

- Seaborn: This is one level-upper version of matplotlib, which was also used in creating visualizations that were more informative and enhanced.

Results and Observations:

- Distribution of sentiments from the analysis

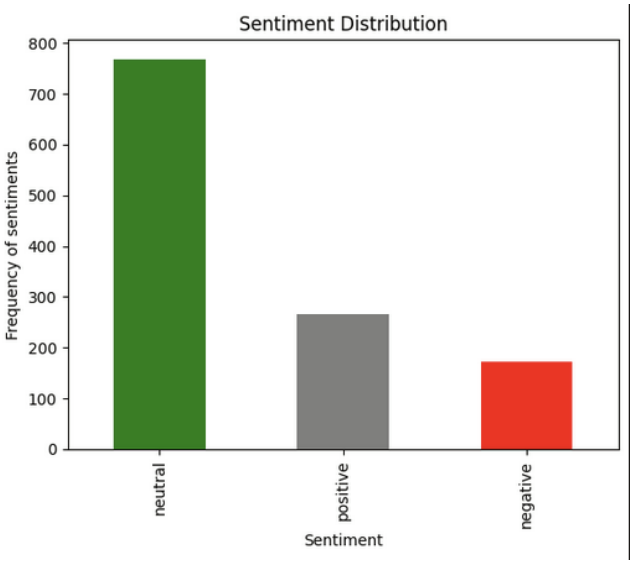

Fig 6.1

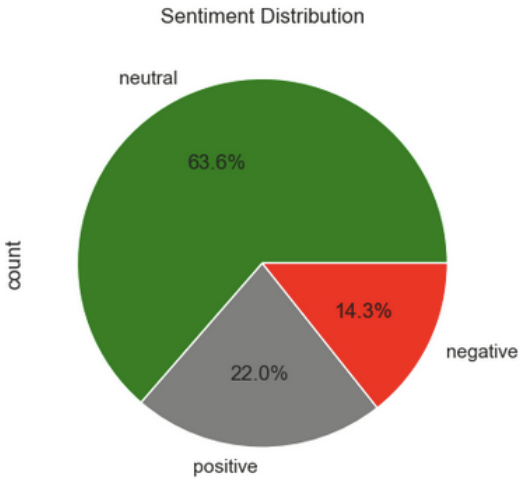

Fig 6.2

- Distribution of Emotions from the analysis

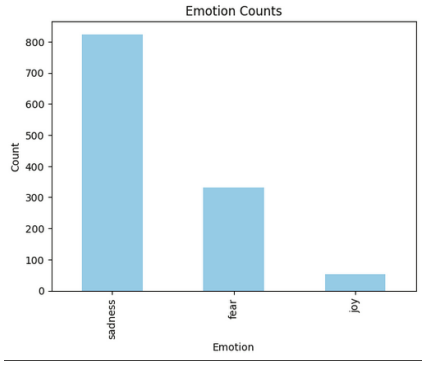

Fig 6.3

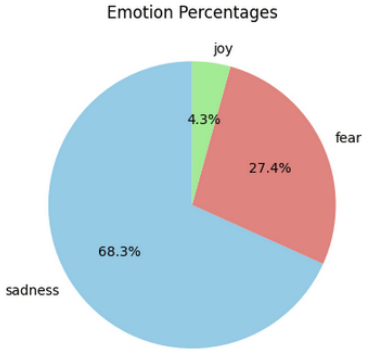

Fig 6.4

- Distribution of Emotion when two nearby sentences are combined

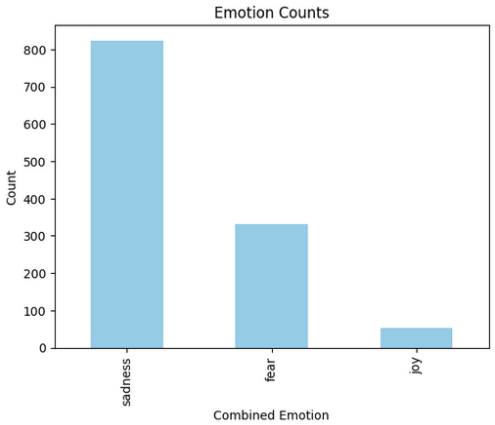

Fig 6.5

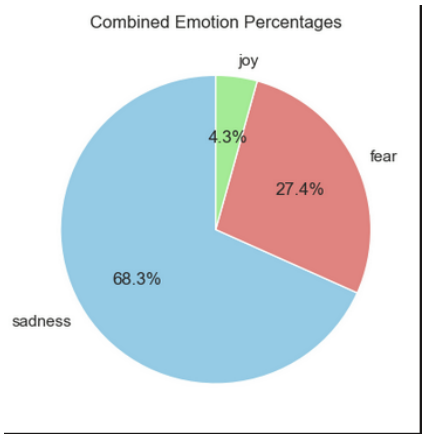

Fig 6.6

- Emotions that differed once neighboring sentences were combined

```
Number of differring emotions of single sentence compared to combined sentencesare:  273
```

Fig 6.7

- Distribution of Sentiment when two nearby sentences are combined

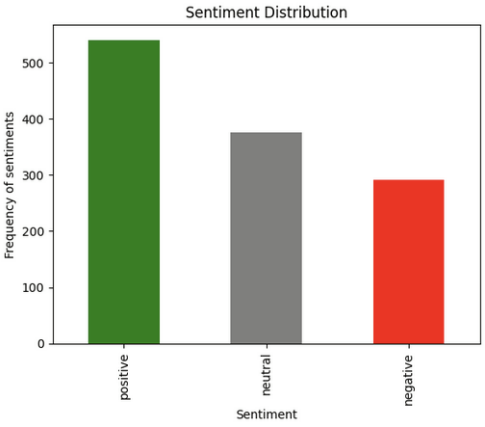

Fig 6.8

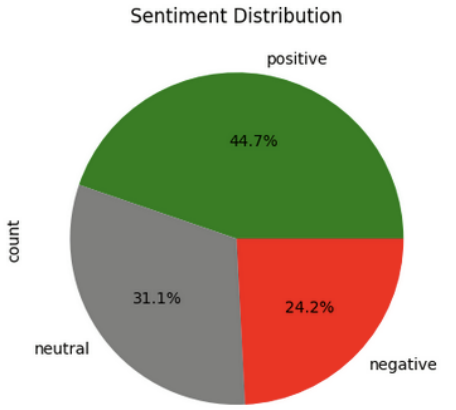

Fig 6.9

- Sentiments that differed once neighboring sentences were combined

```
Number of differring sentiments of single sentence compared to combined sentencesare:  455
```

Fig 6.10

- Correlation between Sentence Length, Sentiment, and Emotions

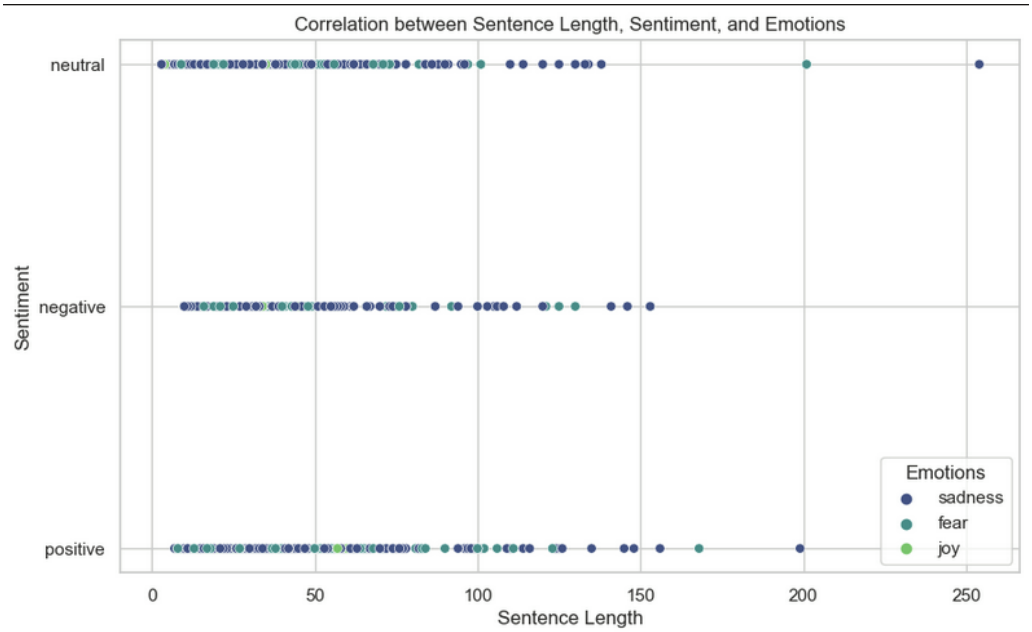

Fig 6.11

- Correlation between Combined Sentence Length, Combined Sentiment, and Combined Emotion

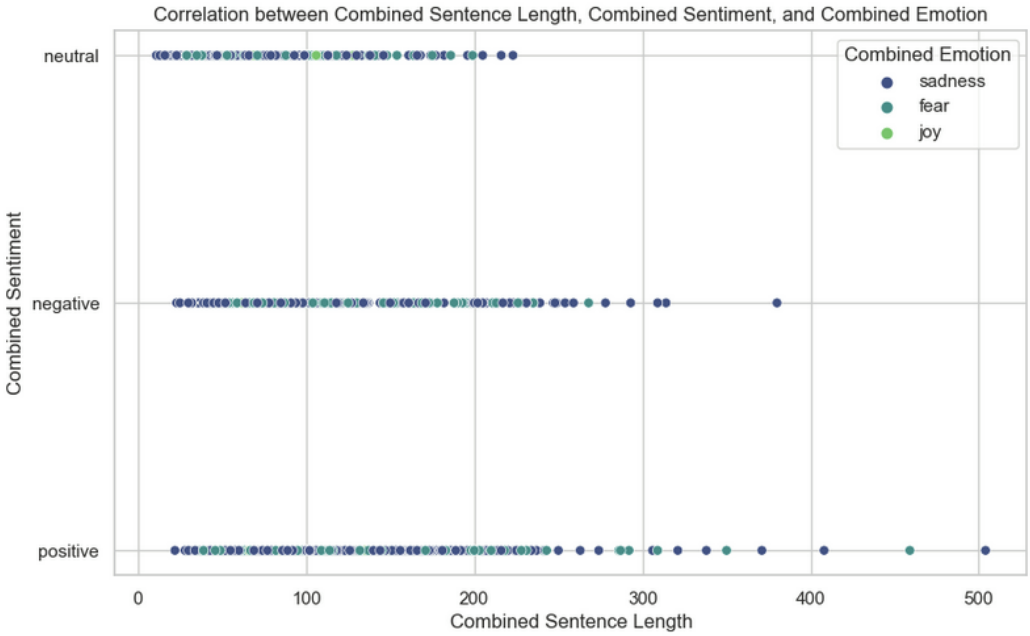

Fig 6.12

- Correlation between sentence length, sentiment, and emotions

	Sentence Length	Sentiment Code	Emotions Code
Sentence Length	1.000000	0.113507	-0.007926
Sentiment Code	0.113507	1.000000	-0.077893
Emotions Code	-0.007926	-0.077893	1.000000

Fig 6.13

- Word cloud of Joy emotion with different sentiments

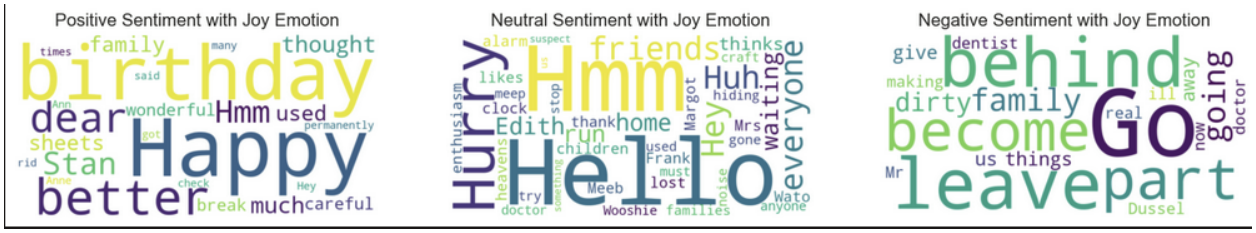

Fig 6.14

- Word cloud of Sadness emotion with different sentiments

Fig 6.15

- Word cloud of Fear emotion with different sentiments

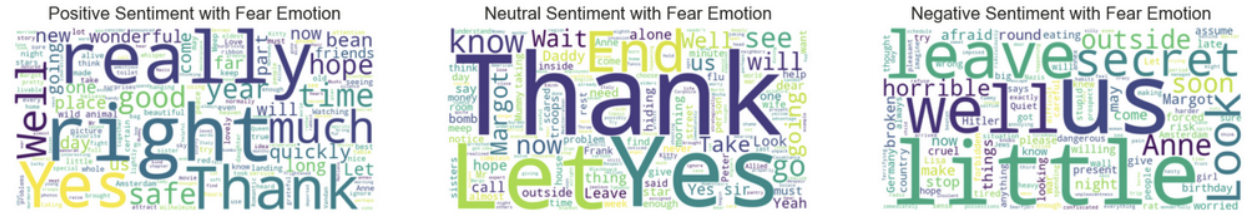

Fig 6.16

- Correlation between Sentiment and Emotions

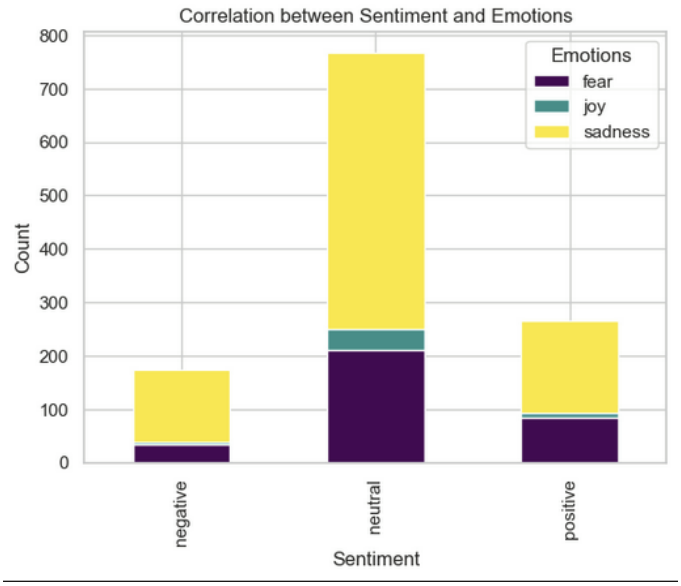

Fig 6.17

- Correlation between Combined Sentiment and Combined Emotion

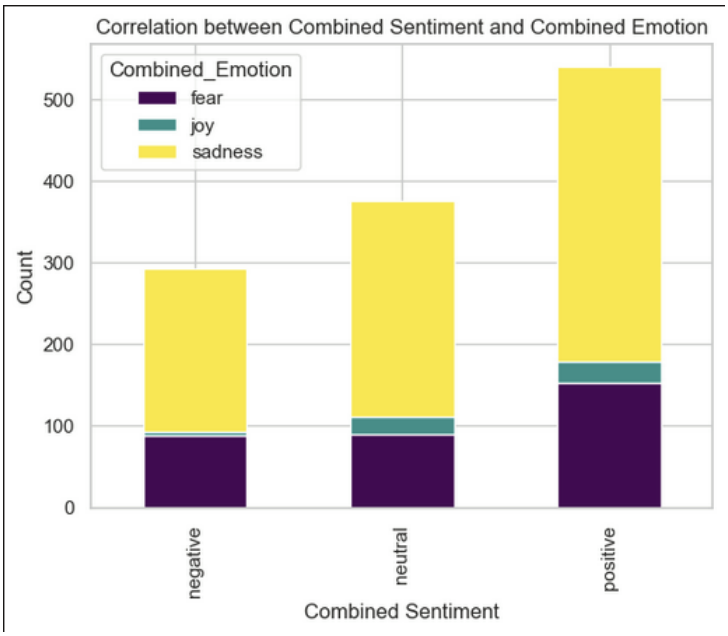

Fig 6.18

- Evolution of emotion throughout the film

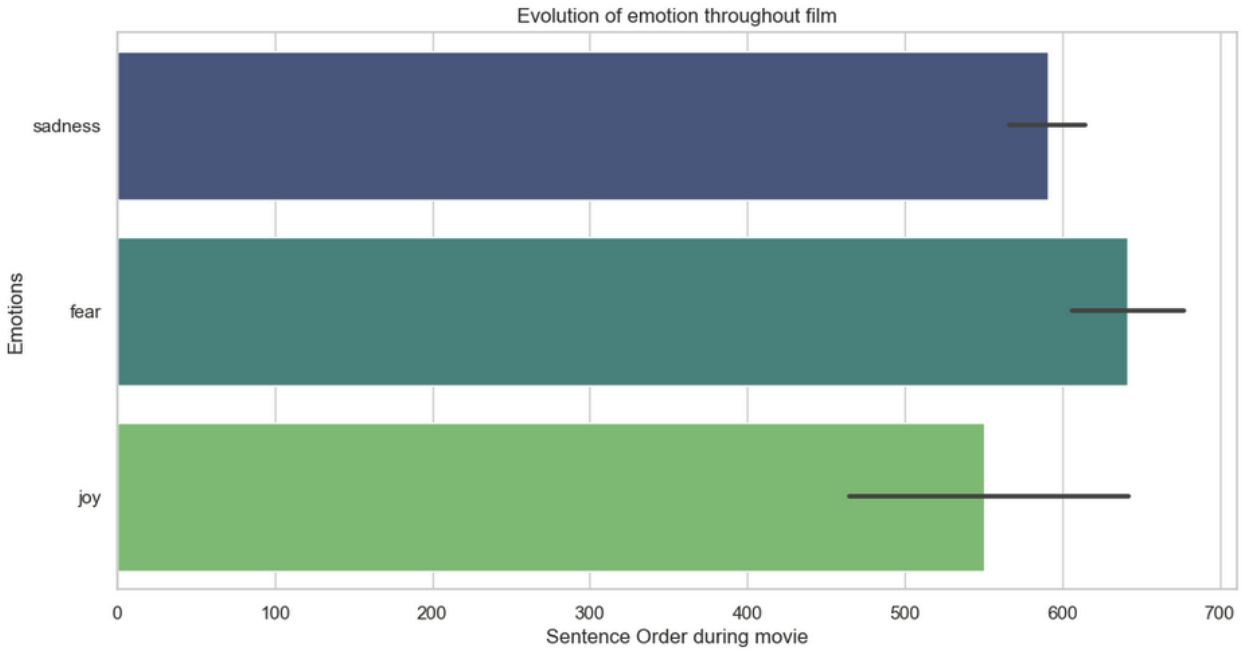

Fig 6.19

Conclusion and Inference:

This project through analysis concludes with answers to the following questions,

- Do sentiments (opinion or view) differ from emotions?
  Observation from Fig 6.1, Fig 6.2, Fig 6.3, and Fig 6.4 determines that sentiments and emotions differ. This again holds then analysis is done for combined sentences which combines neighboring sentences to do analysis. This is proved by Fig 6.5, Fig 6.6, Fig 6.8, and Fig 6.9.

- Are both these the same or different?
  Observation from Fig 6.11, Fig 6.12, Fig 6.17, and Fig 6.18 show that emotions are similarly distributed across all three sentiments [Positive, Negative, Neutral], and further test result from correlation analysis from Fig 6.13 indicates that they are independent of each other.

- How does emotion evolve throughout the movie?
  The movie has exhibited 3 emotions according to the model built. From Fig 6.19, Emotion Joy is prevalent at the start of the movie followed by sadness during the middle phase and fear is concentrated towards the end of the movie.

- Do sentences combined have different emotions and sentiments compared to a single sentence?
  The combination of sentences from the comparison between Fig 6.3, Fig 6.4, and Fig 6.5, Fig 6.6 showed that there was no change in the distribution of the emotions but it did show a difference between emotions displayed for 255 out of 1207 entries of sentences from Fig 6.7. However, in the case of sentiments, from the comparison between Fig 6.1, Fig 6,2, and Fig 6.8, Fig 6.9 showed that there was there was a change in the distribution. There were also 455 out of 1207 entries of combined sentences from Fig 6.10 which showed a different sentiment than single sentences.

Future Work:

- Making Spoken Words Count:
  In the future, the project aims to make our voice-to-text system better by using a bigger set of data to understand emotions in spoken words. The goal is to create a smarter system that recognizes a wider range of feelings, not just in written text but also in how people speak.

- Listening to Emotions in Sounds:
  The analysis shows that our current system might miss some emotions in the way people express themselves through sound. To fix this, future plans include making a model that also understands acoustic emotions and tones instead of just spoken words.

- Seeing Emotions in Faces:
  Emotions aren't only in what we say or how we say it; they also show on our faces. Looking forward, the project aims to build a model that captures facial emotions by analyzing frames. This way, we can get the whole picture of emotions—whether they are in text, sounds, or the way people look.

References and links:

- Papala, G., Ransing, A. and Jain, P., 2023. Sentiment Analysis and Speaker Diarization in Hindi and Marathi Using using Finetuned Whisper: Sentiment Analysis in Hindi and Marathi. *Scalable Computing: Practice and Experience*, *24*(4), pp.835-846.
- Spiller, T.R., Ben-Zion, Z., Korem, N., Harpaz-Rotem, I. and Duek, O., 2023. Efficient and Accurate Transcription in Mental Health Research-A Tutorial on Using Whisper AI for Sound File Transcription.

- Gujjar, J.P. and Kumar, H.P., 2021. Sentiment analysis: Textblob for decision making. *Int. J. Sci. Res. Eng. Trends*, 7(2), pp.1097-1099.
- Diyasa, I.G.S.M., Mandenni, N.M.I.M., Fachrurrozi, M.I., Pradika, S.I., Manab, K.R.N. and Sasmita, N.R., 2021, May. Twitter Sentiment Analysis as an Evaluation and Service Base On Python Textblob. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1125, No. 1, p. 012034). IOP Publishing.
- Faridani, S., 2011, October. Using canonical correlation analysis for generalized sentiment analysis, product recommendation and search. In *Proceedings of the fifth ACM conference on Recommender systems* (pp. 355-358).
- Dos Santos, C. and Gatti, M., 2014, August. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers* (pp. 69-78).
- Duncan, B. and Zhang, Y., 2015, July. Neural networks for sentiment analysis on Twitter. In *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)* (pp. 275-278). IEEE.
- https://en.wikipedia.org/wiki/Whisper_(speech_recognition_system)
- https://openai.com/research/whisper
- https://textblob.readthedocs.io/en/dev/
- https://www.analyticsvidhya.com/blog/2021/10/making-natural-language-processing-easy-with-textblob/
- https://www.questionpro.com/features/correlation-analysis.html#:~:text=What%20is%20correlation%20analysis%3F,the%20change%20in%20the%20other.
- https://wiki.pathmind.com/neural-network
- https://interestingengineering.com/lists/ai-machine-learning-neural-networks