UNIVERSITY OF MALAYA

WQD7023 Data Science Research Project

Multimodal Emotion Detection in Videos using pre-trained LLMs

Guntupu Sowjanya (22085697)

Faculty of Computer Science and Information Technology

Supervisor: DR. Riyaz Ahamed Ariyaluran Habeeb Mohamed

2023/2024

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Guntupu Sowjanya (I.C/Passport No: W2913496)

Matric No: 22085697

Name of Degree: Master of Data Science

Title of Project Paper ("this Work"):

Multimodal Emotion Detection in Videos Using Pre-Trained LLMs

Field of Study: Multimodal Emotion Recognition/Detection

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature: Guntupu Sowjanya Date: 16-08-2024

G.Sowjanya

Subscribed and solemnly declared before,

Witness's Signature  Date: 16-08-2024

Name: Dr. Riyaz Ahamed Ariyaluran Habeeb

Designation: Senior Lecturer

Multimodal Emotion Detection in Videos using pre-trained LLMs

Abstract

The complexity of human emotional expression makes it difficult to accurately detect emotions in videos, a critical component of applications such as affective computing and human-computer interaction. This research project addresses these challenges by suggesting a unique multimodal emotion recognition system using pre-trained Large Language Models (LLMs) in concert with cutting-edge audio and visual processing methods. The research project aims to enable a complete knowledge of the emotional content by means of the integration of textual, audio, and visual data retrieved from videos. The approach highlights the possibility of pre-trained LLMs more especially, LLaMa3 in handling and assessing multimodal data, thereby addressing emotion detection challenges. Comprehensive tests and analysis help to evaluate the system's performance by showing how well it captures and interprets complicated emotional signals. The results of the research project allow to progress emotion detection technology and its possible uses in many fields, including content recommendations, customer service, and individualized education. The project also emphasizes the need of tackling computational complexity, possible biases in pre-trained models, and the requirement of diverse and thorough datasets for next developments in multimodal emotion detection.

Keywords: Emotion detection, Multimodal analysis, Large Language Models, LLaMa3, Emotion Recognition

Acknowledgements

This research project would not have been able to be completed successfully without the help and direction of those around me. For his great insights, support, and unwavering conviction in my talents, my supervisor, Dr. Riyaz Ahamed Ariyaluran Habeeb Mohamed, deserves my most thanks. My research project path has been greatly shaped by his mentoring, which also helps me to overcome obstacles on way. The University of Malaya's Faculty of Computer Science and Information Technology makes me especially appreciative of their tools and chances for doing this research project. At last, I would want to thank my family and friends for their unwavering love and encouragement over this journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

- LLMs: Large Language Models
- AI: Artificial Intelligence
- CRISP-DM: Cross-Industry Standard Process for Data Mining
- CNN: Convolutional Neural Network
- LSTM: Long Short-Term Memory
- RNN: Recurrent Neural Network
- FER: Facial Emotion Recognition
- HRI: Human-Robot Interaction
- NLP: Natural Language Processing
- BERT: Bidirectional Encoder Representations from Transformers
- MFCC: Mel-Frequency Cepstral Coefficients
- PCA: Principal Component Analysis
- SMOTE: Synthetic Minority Over-sampling Technique
- EDA: Exploratory Data Analysis
- STFT: Short-Time Fourier Transform
- ReLU: Rectified Linear Unit
- Tfidf: Term Frequency-Inverse Document Frequency
- ROS: Random Over-Sampling

CHAPTER1: INTRODUCTION

Human communication heavily relies on emotion, which shapes both our interactions with one other and our use of technology. As digital interfaces get more dynamic, improving user experiences in everything from marketing and education to entertainment and treatment depends on the capacity to precisely identify and react to human emotions. Conventional emotion recognition algorithms have mostly analyzed specific data modalities, such textual data, voice patterns, or face expressions. But because emotional expression is by nature multimodal and context-dependent, conventional unimodal methods frequently fall short of capturing the subtlety and complexity of human emotional expression.

Artificial intelligence (AI) has advanced recently, especially in machine learning and neural networks, which has created new opportunities for more complex emotion detection systems. A major advance among these is multimodal emotion detection, which combines several data sources to produce a more complete and precise assessment of emotional states. This project aims to pioneer a system that can more successfully synthesize audio, visual, and textual cues to detect emotions in video content by using the power of pre-trained Large Language Models (LLMs), which have shown remarkable capabilities in processing and producing human-like text.

There are potential and problems when LLMs are included into emotion detection. LLMs are capable of comprehending and producing text with subtle emotional overtones; but novel methods to model training and architecture are needed to extend these models to analyze and incorporate non-textual input. The goal of this work is to investigate and improve LLMs' capacity to interpret multimodal input and use it to generate accurate, real-time predictions regarding human emotions.

1.1 Problem statement

Accurate emotion detection in the videos is one of the most crucial steps for affective computing, human-computer interaction and content recommendation. Even with many advancements, existing systems mostly rely on a single modality of data like textual data/visual

data alone, which many not really help with identifying the human emotions, that are typically identified by a combination of audio, video and textual data.

The study conducted by (Hans & Rao, 2021) used a CNN-LSTM frameworks on facial emotions to demonstrate the useful capabilities in handling the visual data but it doesn't make any use of the textual or audio data by not including them, but they are very useful for the comprehensive emotion detection. In the study of (Mocanu et al., 2023), the authors proposed a novel integration technique of both audio and video data using attention mechanism and deep distance metric learning. Such method is distinguished for its potential in preserving significant data from each type of data, besides enhancing fusion of features from different modalities. Through their approach, it has been possible to establish notable working rates of the system in detecting emotions through Multimodal Emotion recognition, especially in datasets such as RAVDESS and CREMA-D.

Additionally, the study by (Aslam et al., 2023) illustrates that, despite recent multimodal sentiment analysis techniques for integrating audio, visual, and text data, there still remain challenges. Multimodal systems effectiveness entails the use of advanced models to deal with the problem of data source heterogeneity, the problem of alignment and synchronisation of different modalities, and the extraction of discriminative features from each modality to enhance prediction accuracy. (Heredia et al., 2022) investigated an adaptive multimodal emotion detection architecture intended especially for social robots. The versatility and robustness of multimodal techniques are demonstrated by this architecture, which assesses various combinations of text, audio, and facial input to classify feelings. Even in the lack of one or more data kinds, the study emphasizes the need of combining several modalities to attain competitive performance.

Furthermore, the significant contribution by (Bhattacharya et al., 2021) points out on how there is a need for advancement in the multimodal emotion detection systems to manage the increasing complexity of the online interactions and the need for personalization of the social media. Research by (Middya et al., 2022) has demonstrated that despite the potential of deep learning models to identify people's emotions the actual efficiency of these models is often limited by the used architectures and datasets. For example, when an audio-visual model is employed to learn audio-visual representations, the model must be very precise in extracting the relevant information as well as of the correlations between such modalities.

In this research, the aim is to bridge these research gaps by creating an innovative multimodal emotion detection method that makes use of the benefits of pre-trained LLMs, as well as advanced audio and visual processing techniques. The objective is to utilize and combine the data from different sources like textual, audio and visual data in order to gain a better understanding of the emotions in the videos. This could help with the enhancement of the accuracy and usefulness of the emotion detection techniques in the real-time situations. This approach can help with the improvement of the technological robustness of the emotion detection methods and the ability to successfully implement these developments by addressing the practical issues.

1.2 Research Questions

Q1) How do pre-trained Large Language Models (LLMs) perform in emotion recognition when audio, visual, and textual data are integrated?

Q2) How can pre-trained Large Language Models (LLMs) be used to integrate the textual, visual and audio data to enhance emotion detection?

Q3) How could comprehensive assessments and evaluations in controlled environments be used to assess the reliability and accuracy of a multimodal emotion detection model?

1.3 Research Objectives

Objective 1: To investigate the existing multimodal for emotion detection while integrating textual, visual and audio data which can enhance emotions are detected in videos.

Objective 2: To utilize a pre-trained Large Language Model (LLM) based framework for the multimodal data (textual, visual and audio) for improving the emotion detection

Objective 3: To evaluate the multimodal emotion detection system's performance by comprehensive evaluation and analysis.

1.4 Research Significance

This research project is significant as it combines text, audio, and visual data to enhance emotion identification in videos. A more accurate system will result from this method than from employing only one kind of data. Better emotion recognition can enhance the ease and interest of interactions with technology, such as social robots and virtual assistants. Helping us comprehend how various data types interact, also helps with affective computing. Personalized education, customer service, and content recommendations are just a few of the useful applications for the data. In addition, this research will result in a more dependable and efficient system for practical application, promoting creativity in domains such as computer vision, natural language processing and machine learning.

CHAPTER 2: LITERATURE REVIEW

The field of emotion detection has made great progress, notably in the analysis of video data, which provides a wide range of emotional clues via visual, aural, and textual modalities. Investigating these several senses has produced advanced methods meant to catch the minute subtleties of human emotional expression. Focusing on three main areas emotion recognition in videos, multimodal emotion detection, and the developing subject of emotion detection utilizing pre-trained Large Language Models (LLMs) the component of this research project on the literature review explores the body of work already in existence in emotion detection. The study looks at the advantages and drawbacks of several strategies, therefore stressing the possibilities and difficulties in this interesting field. It also emphasizes the need of combining several modalities to get a more thorough and precise understanding of emotions in videos.

2.1 Emotion Detection in Videos

Article 1: A CNN-LSTM based Deep Neural Networks for Facial Emotion Detection in Videos (Hans & Rao, 2021)

(Hans & Rao, 2021) proposed a model that uses CNN with LSTM to detect facial emotion of people in video sequences. Training this model they utilized the CREMA-D dataset and evaluated the model on the RAVDEES dataset, and the accuracy of the model is 78.52% and 63.35%, respectively. This research paper also focuses on the use of CNNs for feature extraction as well as on the use of LSTMs for modelling the temporal variations of facial expressions.

Article 2: Deep Emotion Change Detection via Facial Expression Analysis (Han et al., 2023)

(Han et al., 2023) proposed a method of training a weak supervisor deep learning system based on static images of the human face for identifying changes in the emotional state. This strategy also helps capture vaguer changes in actors' emotion without relying on videos with many annotations to validate strong performance in different datasets. This is especially the case in dynamic applications such as social robots and interactive media making this a much

superior method over traditional emotion recognition as the latter does not take temporal factors into consideration.

Article 3: Deep Facial Emotion Recognition Using Local Features Based on Facial Landmarks for Security System (An et al., 2023)

(An et al., 2023) elaborated boosting the FER for security applications and analyzing local features derived from facial landmarks, eye distance, brow distance, nose distance, mouth distance. They based their approach on Euclidean distances and ENS classifiers; they gained roughly 25 percent accuracy compared with more conventional algorithms for object recognition and demonstrated high sensitivity to shifts in lighting and background. This study contributes greatly to emotion recognition particularly to the identification of important emotions such as fear and anger important in security systems.

Article 4: Detection of Human Emotions Through Facial Expressions Using Hybrid Convolutional Neural Network-Recurrent Neural Network Algorithm (Manalu & Rifai, 2024)

In their study of face expression identification with a hybrid CNN-RNN model, (Manalu & Rifai, 2024) improve the ability to identify subtle human emotions from video data. For their study, they utilized the Emotional Wearable Dataset 2020, which contains a variety of hitherto undiscovered emotions including excitement and amusement. Three models—MobileNetV2-RNN, InceptionV3-RNN, and a bespoke CNN-RNN—are compared in this work; the latter achieves significant accuracy gains. This work shows great promise for practical uses like improving interactive digital communication and is an excellent example of the increasing sophistication in emotion detecting technologies.

Article 5: Emotion recognition at a distance: The robustness of machine learning based on hand-crafted facial features vs deep learning models (Bisogni et al., 2023)

(Bisogni et al., 2023) described a study on the comparison of machine learning algorithms and deep learning techniques for emotion detection based on handcrafted facial features. They checked the work of deep learning on the Extended Cohn-Kanade dataset and found out that while it works best on high-quality images, it is quite poor in low-quality images. The study also described that the robustness of the handcrafted feature-based models maintained the

similar trend because these models are more usable in controlled and real-world surveillance environment where input might not be so clean.

Article 6: Facial Emotion Recognition for Photo and Video Surveillance Based on Machine Learning and Visual Analytics (Kalyta et al., 2023)

(Kalyta et al., 2023) examined FER in surveillance video with attention brought to change detection in the Expression using low-resolution images via a video camera. Specifically, they used a new approach based on the geometry of faces and hyperplanes to obtain better interpreting results and model performance. Their endeavor shows better recognition of the changes in human faces and provides a better security system to restrict unwanted elements in areas like airports, commercial complexes, etc.

Article 7: Facial Emotion Recognition Using Transfer Learning in the Deep CNN (Akhand et al., 2021)

(Akhand et al., 2021) proposed effective transfer learning in deep Convolutional Neural Networks (CNNs) to improve the FER. They trained and tested original and transferred models such as VGG-16, ResNet-50 using facial emotion dataset such as KDEF and JAFFE and get better accuracy over conventional methods. This method stands out for the identification of various affective states from different perspectives, which proves ideal for real-time, accurate emotion assessment.

2.2 Multimodal Emotion Detection

Article 8: Attention-based Multimodal Sentiment Analysis and Emotion Recognition Using Deep Neural Networks (Aslam et al., 2023)

(Aslam et al., 2023) proposed the "Attention-based Multimodal Sentiment Analysis and Emotion Recognition (AMSAER)" system that utilizes deep learning to provide the combined solution for text, audio, and visuals. Thus, improving attention mechanisms in the AMSAER

model contributes to increasing the accuracy of sentiment and emotion classification and, therefore, identifying the once-hidden features in the joined modalities. A comparison with other approaches on the IEMOCAP dataset shows that their tests were more effective, which proves the added benefit of this progressive method in sentiment analysis.

Article 9: Adaptive Multimodal Emotion Detection Architecture for Social Robots (Heredia et al., 2022)

(Heredia et al., 2022) proposed adaptive multimodal emotion detection architecture for social robots in which HRI had improved. This makes the system a combination of text, audio, and visual data, within which the robots can read people's emotions and respond to them in the most precise manner possible, in real-time. One of the important aspects is the EmbraceNet+ fusion approach that helps to optimize the process of combining various types of input data and enhance the convergence of emotion recognition even if some of the inputs are missing or of poor quality.

Article 10: Deep Learning Based Multimodal Emotion Recognition Using Model-Level Fusion of Audio–Visual Modalities (Middya et al., 2022)

(Middya et al., 2022) proposed a study on model-level fusion specifically for emotion recognition using parallel feature extractor network for both audio and video streams. Their work extended the co-integration of these modalities thereby improving the accuracy of the experiments on public datasets RAVDESS and SAVEE. The findings expose deep learning's ability to handle vast data and corroborate attaching multiple modalities for the enhancement of recognizing emotions.

Article 11: Exploring the Contextual Factors Affecting Multimodal Emotion Recognition in Videos (Bhattacharya et al., 2021)

(Bhattacharya et al., 2021) showed where and with whom influences the perception of emotion in the video by scraping 2,176 YouTube videos for analysis. They also concluded that while using features extracted from more than one modality outperforms single mods and two mods, the performance is best when using only video modality in male speakers and short videos. The analysis shows that gender and expression duration correlate with the accuracy of

recognition and emphasizes the impact of these ideas on creating new intelligent technologies for various fields.

Article 12: Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning (Mocanu et al., 2023)

(Mocanu et al., 2023) presented an efficient cross-modality multimodal audio-video fusion framework for the MEmotion recognition using attention mechanisms and Deep Metric Learning exists. Their scheme uses 3D-CNN for the VIDEO data with spatial, channel, and temporal attention and 2D-CNN for the AUDIO data with temporal attention only. This configuration nicely handles complex relationships between modalities, gets better feature fusion and discriminative performance. This model appeared to be doing well in both RAVDESS and CREMA-D datasets which suggests the fact that this proposed model is quite useful in instances where fine-grained emotion analysis is needed.

2.3 Emotion Detection in Videos Using Pre-trained LLMs

Article 13: Emotion Detection for Social Robots Based on NLP Transformers and an Emotion Ontology (Graterol et al., 2021)

To improve emotion identification in social robots, (Graterol et al., 2021) provided a novel framework that integrates an emotion ontology (EMONTO) with NLP transformer. With this method, robot-human interactions can be enhanced by robots' precise interpretation of human emotions from textual and speech data. Tested in settings including museums, the system competes with top transformer-based models by efficiently classifying emotions to customise robot reactions. One major step towards creating interactive and sympathetic robots is the combination of semantic technology with natural language processing.

Article 14: Capacity of Generative AI to Interpret Human Emotions from Visual and Textual Data: Pilot Evaluation Study (Elyoseph et al., 2023)

(Elyoseph et al., 2023) examined ChatGPT-4 and Google Bard to see how well generative AI models can grasp text and images to predict human emotions. It's interesting that ChatGPT-

4 can recognize these characteristics in textual and explicit emotional components like human performance on the RMET and LEAS tests. Google Bard is better at recognizing emotions from text than other AI models, but it cannot read visual cues or recommend development areas. This study emphasizes exploring common difficulties and AI's growing capabilities in mental health applications.

Article 15: Detecting Emotions Behind the Screen (Alkaabi et al., 2022)

(Alkaabi et al., 2022) used BERT transfer learning to detect emotions in educational platforms. Their work shows students using BERT to analyze text messages on an instructional app to determine emotional states like tension or concern. Our technique outperforms standard machine learning at 91% accuracy. The study shows how BERT and the NRC lexicon can improve emotion identification by disclosing students' key negative emotions and expressive phrases. The COVID-19 pandemic-related shift to online learning makes this approach promising for mental health experts to design focused therapies.

Conclusion

The following points presents key research to emphasize the importance of emotion detection where deep learning and NLP should increase recognition performance significantly. Use of multiple data types, that is, sound, video and text data has enhanced system reliability in a very big way. Various studies based on the use of attention mechanisms and model fusion have given promising results in the processing of the fine-grained emotion signals (Aslam et al., 2023); (Middya et al., 2022); (Mocanu et al., 2023). The up use of generative AI models in mental health as well as in the education sector has shown possibility for general implementation (Elyoseph et al., 2023); (Alkaabi et al., 2022). Nevertheless, the following issues present some of the concerns for further development: dataset diversification, or the creation of technology that identifies other emotional outlooks besides the happiness-sadness binary; and ethical problems that surround emotion recognition technology. Making the data more comparable and integrating it should be the focus of future research endeavors, and the ethical considerations of using such technologies should be studied and understood.

CHAPTER 3: RESEARCH METHODOLOGY

The CRISP-DM framework, which is an organized approach that directs the data mining process from business knowledge to deployment, is used in this project's study technique. Aiming to use the advanced models in integrating and analyzing text, video, and audio data, the project detects multimodal emotions in videos using pre-trained Large Language Models (LLMs). The approach consists of a thorough preprocessing stage where data from several modalities are ready, and features are retrieved using the MELD dataset. The pre-trained LLMs then evaluate these traits for emotional recognition capacity. The method explores the performance of these cutting-edge models in a multimodal environment and reveals their possibilities in managing challenging emotional signals found in video data.

3.1 Research Design

The primary objective of the research design is to create a multimodal emotion detection system that utilizes pre-trained Large Language Models (LLMs) to integrate audio, video, and text data. The first steps of the procedure involve gathering and analysing data from the MELD dataset. A quality check is then performed to correct any discrepancies or gaps in the data.

Data is cleansed, problems with audio or video sections are fixed, and appropriate synchronizing across all modalities is guaranteed in the data preparation stage. Tools including Librosa for audio, OpenCV for video, and BERT for text are used in feature extraction; subsequently, enhancement and integration results in an organized dataset.

Pre-trained LLMs and other models catered to each modality are applied for the modelling stage. To evaluate their performance, the models are scored using criteria including accuracy, precision, recall, F1 score, ROC AUC. Using tools like Streamlit, the system is finally implemented in real-time under constant monitoring to preserve performance and control changes.

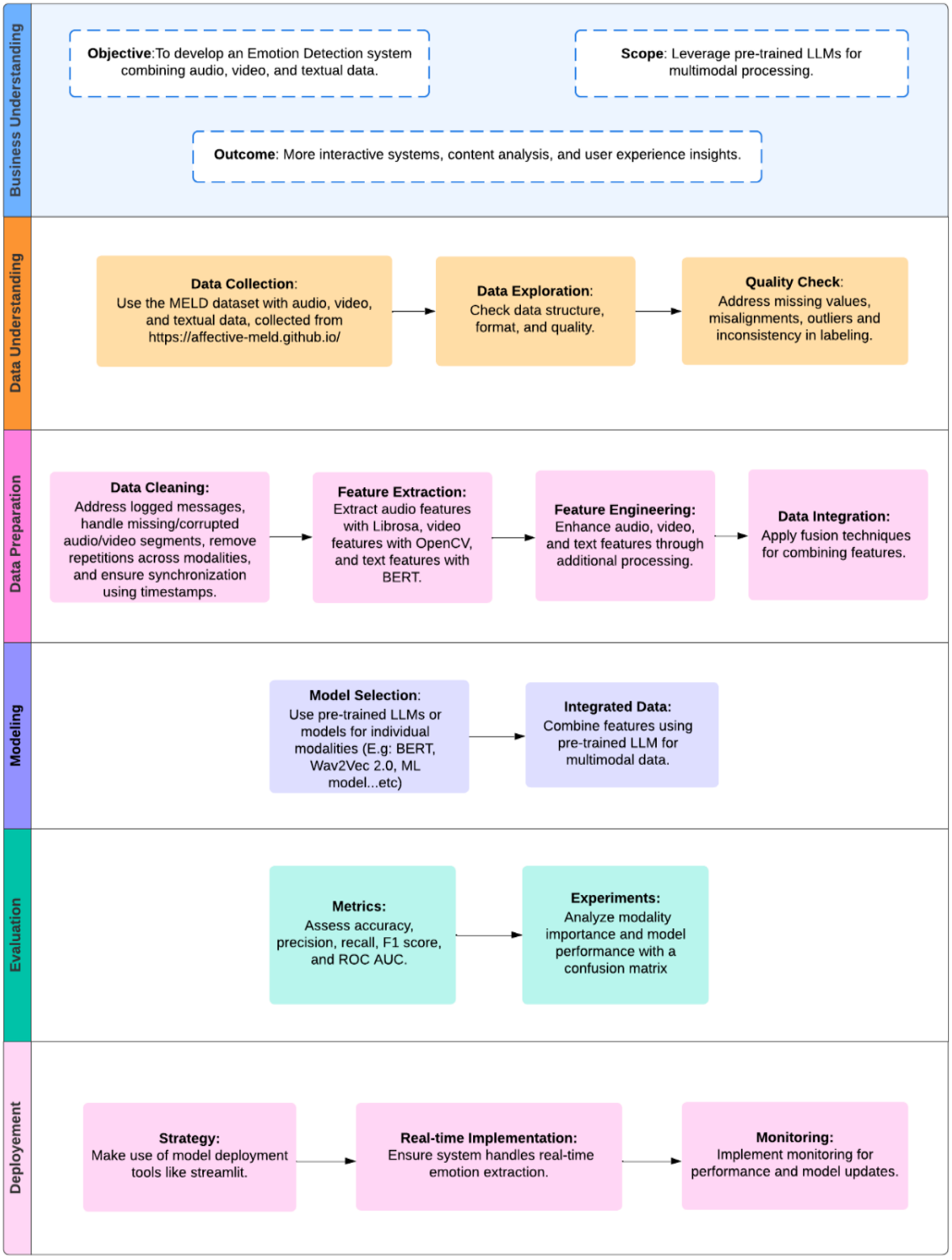

Fig 3.1 Research Design

3.2 Business Understanding

Objective

To develop an Emotion Detection system having strong foundations in audio, video, and textual data for detailed and precise Emotion Analysis.

Scope

Using the abilities of pre-trained LLMs in terms of multimodal processing and combining them to use the capabilities of the model to comprehend complex emotional cues in messages sent across different communication channels.

Outcome

The development of more interactive systems, content analysis, and user experience studies providing in-depth views into emotional states.

3.3 Data Understanding

Data Collection

The data to be used for the analysis is the MELD dataset (Poria et al., 2018), (Chen et al., 2018), annotated with synchronized audio, video, and textual data containing emotional states.

Data Exploration

Preliminary checking of data to get an idea about the structure, format, and quality of the dataset in all of three modalities.

Quality check

The process may be undertaken to single out some wrong data or integrity issues, such as missing values, misalignments in the data-synchronization, and inconsistency in labelling in the dataset.

3.3.1 Exploratory Data Analysis (EDA)

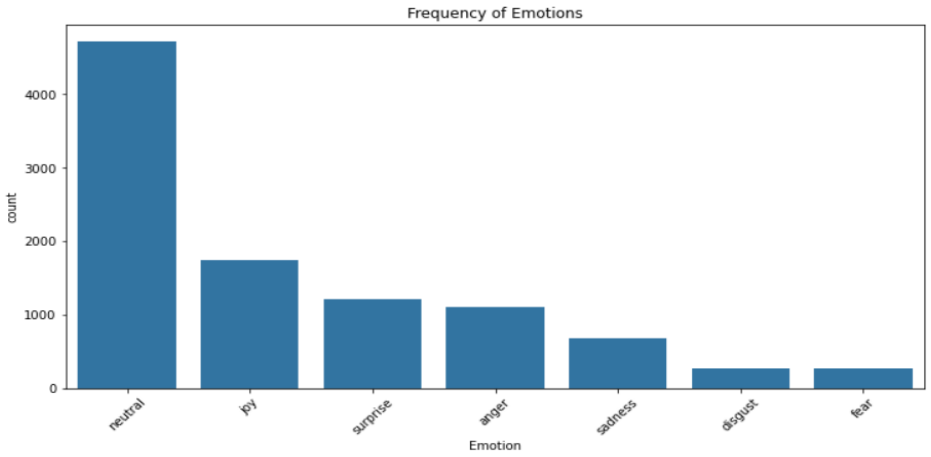

Fig 3.2 Frequency of Emotions

Frequency of Emotions

- "Neutral" is the most common emotion, followed by "joy", "surprise", and "anger".
- "Sadness", "disgust", and "fear" are the least frequent, indicating an **imbalance** in the dataset.

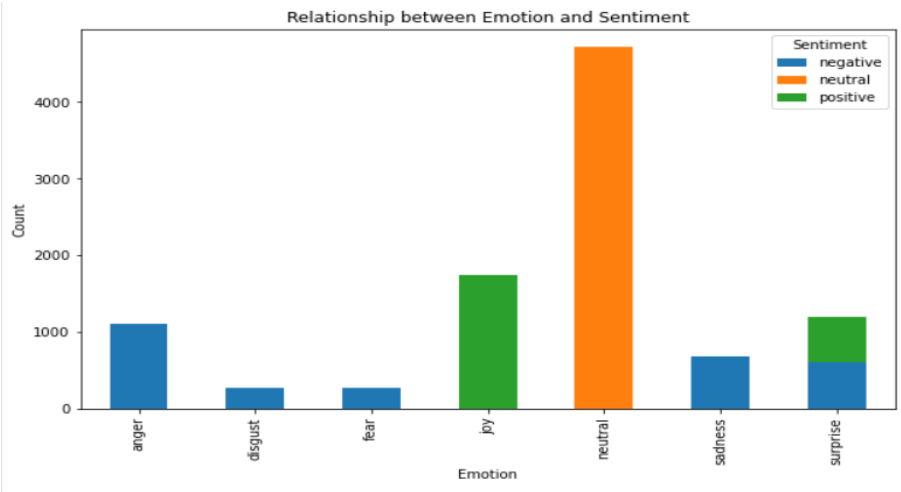

Fig 3.3 Relationship between Emotion and Sentiment

Relationship between Emotion and Sentiment

- "Neutral" sentiment dominates, especially in "Neutral emotion".
- "Joy" is mostly "positive", while "Anger", "Disgust", and "Sadness" are "negative". "Surprise" has both "positive" and "neutral" sentiments.

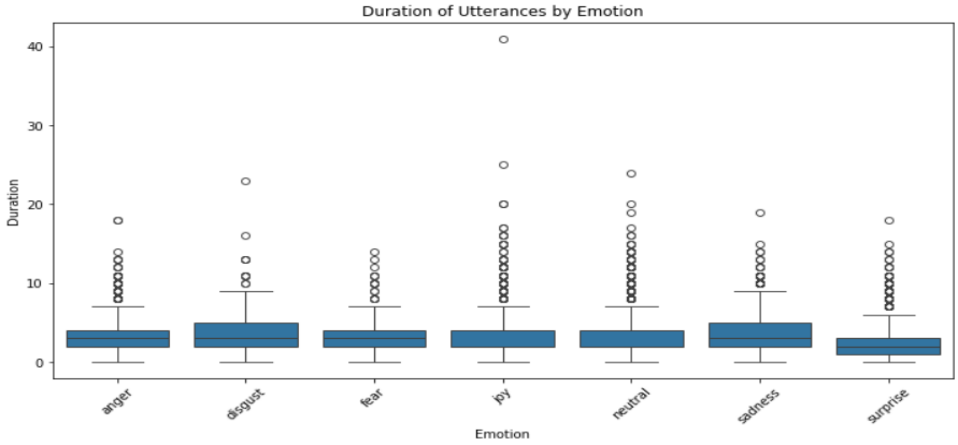

Fig 3.4 Duration of Utterances by Emotion

Duration of Utterances by Emotion

- Most utterances are under 10 seconds for all emotions, with some outliers.
- Median durations are similar across emotions, indicating consistency in utterance length.

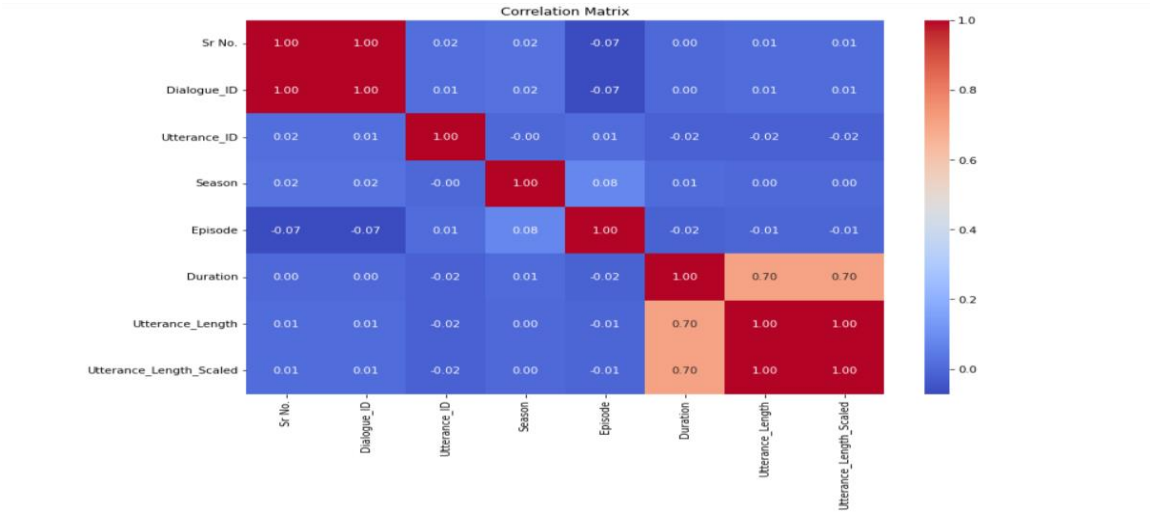

Fig 3.5 Correlation Matrix

Correlation Matrix

- High correlation between Utterance_Length and Utterance_Length_Scaled.
- Moderate correlation between Duration and Utterance_Length; other features show low correlations, indicating independence.

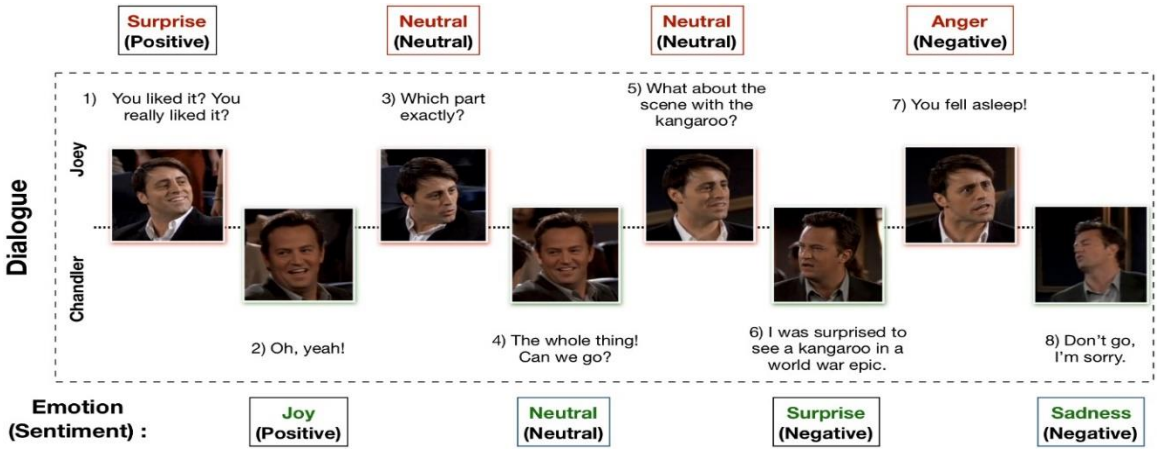

Fig 3.6 Dialogue vs Emotion

Dataset statistics

Statistics	Train	Dev	Test
# of modality	{a,v,t}	{a,v,t}	{a,v,t}
# of unique words	10,643	2,384	4,361
Avg. utterance length	8.03	7.99	8.28
Max. utterance length	69	37	45
Avg. # of emotions per dialogue	3.30	3.35	3.24
# of dialogues	1039	114	280
# of utterances	9989	1109	2610
# of speakers	260	47	100
# of emotion shift	4003	427	1003
Avg. duration of an utterance	3.59s	3.59s	3.58s

Fig 3.7 Dataset Statistics

3.4 Data Preparation

3.4.1 Text Data Preparation

- Tokenizing the text data into separate words and punctuation marks.
- Eliminating stop words common English terms like "the", "and" etc. that, in the context of emotional detection have little bearing.
- For uniformity, lowercase all text data.
- Lemmatizing words to their base or dictionary form that is, "running" becomes "run."

3.4.2 Audio Data Preparation

Every audio recording underwent preprocessing with

- Using Short-Time Fourier Transform (STFT), translate the audio file into a time-frequency model.
- From every STFT frame, obtaining 128-dimension Mel-Frequency Cepstral Coefficients (MFCCs).
- Zero mean and unit variance MFCC normalizing process

Removal of outlier

PCA was used to lower the dimensionality of the pre-processed data hence removing outliers from the dataset. Retaining the top twenty main components allowed one to account for around 95% of the total variance. Outliers and eliminated from the dataset were the top 5% of samples having the highest Euclidean distances from the mean.

Data Balancing

With a notable percentage of samples categorized as "neutral," (60%), the remaining dataset was skewed. New minority class samples were synthesized using Random Over-Sampling (ROS) to handle this problem. Synthetic sample counts were kept under control such that the dataset's majority and minority classes had comparable representation.

3.4.3 Text and Audio Data Preparation

Using statistical techniques including mean and standard deviation, the pre-processed data was then examined for any anomalies or outliers. Any records with values more than two standard deviations away from the mean were deleted, therefore producing a dataset around 9,500 records. (positive, neutral and negative emotions).

Outlier Removal using PCA

Principal Component Analysis (PCA) was used to preprocess the data and eliminate any last outliers, hence lowering the dimensionality of the data. Retaining much of the original information, PCA is an unsupervised technique projecting high-dimensional data onto a lower-dimensional space. This work uses PCA to lower the dimensionality of the audio features, therefore producing a dataset with less dimensions and less noise.

SMOTE based dataset balancing

After the outliers were removed and the class distribution was slanted toward one class negative emotions the pre-processed sample was imbalanced. Synthetic Minority Over-sampling Technique (SMOTE) was applied to create fresh synthetic samples for the minority class positive emotions to handle this problem. Popular oversampling method SMOTE generates new samples by interpolation between current minority class samples.

3.5 Modelling

The project focuses on using pre-trained Large Language Models (LLMs) to the challenge of multimodal emotion identification in videos. This is combining textual, video, and audio data into an integrated model capable of precisely interpreting and categorizing emotional cues. The method uses the pre-trained on enormous volumes of data capabilities of LLMs to manage complicated interactions across several modalities. The modeling method consists in choosing suitable architectures, optimizing the models to the job of emotional detection, and assessing their performance to guarantee that the combined model efficiently captures the nuances of multimodal emotional expression.

3.5.1 Text Model Overview

Processing and evaluating textual input for emotional identification depends critically on the Text Model. The main objective is to arrange unstructured text data such that it is fit for machine learning models more especially, the LSTM model in this case.

Text processing Class

Function: Preprocessing of the input text data falls to this class. Tokenizing, removing stop words, lemmatizing, and translating the text into a numerical representation the LSTM model can grasp comprise a few processes here.

Method

- preprocess_text(text: str) returns a list. Breaking the input text string down into tokens, eliminating extraneous characters, and standardizing the structure, this approach cleans and processes the content. The result is a ready for embedding list of tokens.
- Ensuring that the text data is clean, consistent, and in a format allowing the LSTM model to efficiently learn patterns and relationships inside the text depends on the preprocessing stage. This class basically prepares the text data, therefore enabling the collection of the semantic and syntactic subtleties required for correct emotional prediction.

Data Preprocessing Class:

Function: This class tokenizes and pads sequences to guarantee consistent input size, therefore handling the last preparation of text data for the LSTM model.

Methods:

- **texts_to_sequences():** Creates integer sequences from preprocessed text, where each integer denotes a token word in the text.
- **pad_sequences() :** Pads these sequences such that, for batch processing in LSTM, their lengths all match.

This class helps the LSTM model to batch process input data by converting and padding text sequences, hence enabling effective training and inference.

Model Overview for LSTM

Designed especially to manage sequential data, such text, the Long Short-Term Memory (LSTM) model is a variant of recurrent neural network (RNN). It is especially good for jobs like sentiment analysis and emotion recognition, where the meaning of words greatly depends on their order; it also excels in learning from time-dependent data.

LSTM Model Class

Architecture

- **Embedding Layer:** Captures semantic meaning via dense vector of fixed size converting every word (token) in the text. This layer creates a more expressive and lower-dimensional space from the integer input sequences.
- **LSTM Layer:** This is the model's central component. It analyzes the contained sequences, over time capturing linkages and relationships. Understanding context in text depends on the LSTM layer remembering significant information over lengthy sequences.
- **Dropout Layer:** During training, a random fraction of input units are set to zero in order to prevent overfitting. This increases the model's generalizable strength and resilience to fresh data.
- **Layer of Fully Connected (FC) nature:** The last layer producing the projected emotional category transfers the LSTM layer's output to the emotion labels with a softmax activation function.

PyTorch Training Class

Training Process

- **Loss Function:** CrossEntropyLoss is commonly employed for multi-class emotion prediction, in which the model gains knowledge by reducing the discrepancy between the labels that are predicted and those that are actual.

- **Optimizer:** Adam is often used for its efficiency in managing complicated models like LSTM and vast amounts.
- **Forward Pass:** The fully connected, LSTM, and embedding layers all get the text data. To find the loss, one compares the model's forecasts with real labels.
- **Backward Pass:** The optimizer changes the model weights to reduce the loss, hence enhancing the model's performance over time.

The capacity of the LSTM model to preserve a recall of past inputs makes it quite appropriate for jobs requiring comprehending sequences, such emotion recognition from text. Together with suitable preprocessing and training techniques, the architecture lets the model efficiently capture the subtleties of human emotions as stated in text.

The diagram demonstrates the interaction between components and the flow of data, emphasizing the preprocessing, modelling, and training of the data to generate a predictive text-based model.

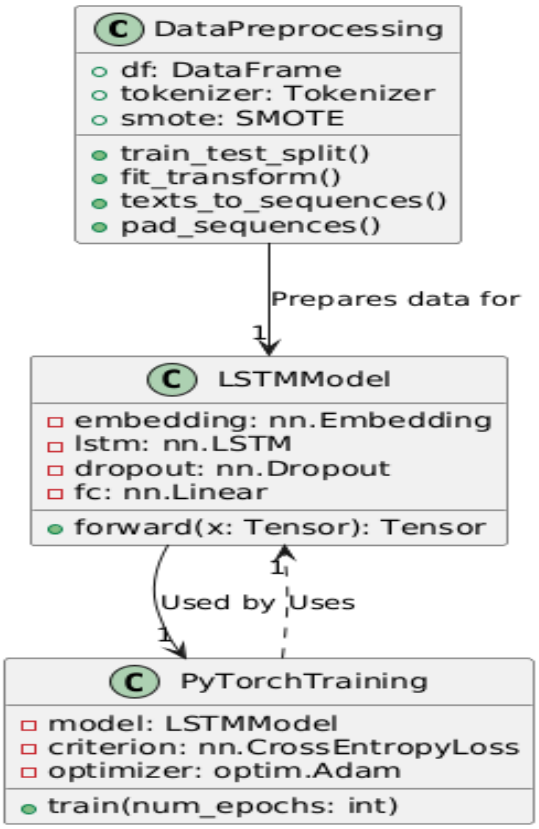

Fig 3.8 UML Diagram for Text

3.5.2 Audio Feature Extraction and Analysis

1. Audio Feature Extraction Using MFCC

Processing audio data for machine learning applications including emotion detection depends primarily on audio feature extraction. The main objective is to convert unprocessed audio inputs into a set of features capturing the fundamental audio properties. The Mel-Frequency Cepstral Coefficients (MFCC) are among the most often applied methods for this aim.

MFCC Extraction

Mel-frequency cepstral coefficients (MFCC)

Often employed in speech and audio processing, MFCCs are a representation of a sound's short-term power spectrum. Crucially for varied emotional tones, they catch the timbral texture of audio.

Procedures

- Pre-emphasis: To boost the higher frequencies, the audio stream is first pre-emphasized. This stage helps to balance the spectrum and raise the signal to noise ratio.
- Framing: To capture the temporal properties of the sound, the continuous audio input is split into overlapping frames. Usually, every frame runs 20 to 40 milliseconds.
- Windowing: To reduce signal discontinuities at the start and finish of each frame, a Hamming window is applied to each one.
- Fast Fourier Transform (FFT): The FFT provides a spectrum for each windowed frame by transforming them from the time domain to the frequency domain.
- Mel-Scale Filtering: The spectrum is run through a filter bank that uses the Mel scale, which replicates the way that humans perceive sound, to scale frequencies.
- Logarithm: To highlight how loud something is perceived; the power spectrum is transformed to a logarithmic scale.
- Discrete Cosine Transform (DCT): The audio signal is finally represented in a compact manner by the MFCCs, which are created by applying the DCT to the logarithm of the power spectrum.

MFCCs are perfect for jobs like emotion recognition in speech since they are rather good in catching the subtleties of audio signals and enable one to differentiate between little variations in sound.

2. PCA Analysis on Audio Features

Overview: A dimensionality reduction method called principal component analysis (PCA) lowers the feature count while yet maintaining as much variance as feasible. PCA is used to the MFCC features in the framework of audio feature extraction to lower the dimensionality, so enabling more manageable data for next processing phases.

PCA Process

PCA aims primarily to preserve the most important variance in the data by converting the high-dimensional feature space into a lower-dimensional space. This lowering of dimensionality's course helps to accelerate the training process.

Procedure

- Standardization: To start, the MFCC characteristics are set to a mean of zero and a variance of one. PCA is sensitive to the magnitude of the features, hence this stage is important.
- Covariance Matrix Computation: To determine how several features differ from one another, a covariance matrix is generated.
- Eigen Decomposition: The covariance matrix's eigenvalues and eigenvectors are calculated. The eigenvalues show the degree of variation along the lines of maximal variance; the eigenvectors define those directions themselves.
- Feature Transformation: The data is transformed into a new feature space with less dimensions by projecting the original features onto the eigenvectors.
- variation Retention: To make sure that most of the variation in the original data is preserved in the reduced feature set, the highest principal components are usually chosen based on the total explained variance.

Eliminating duplicate information helps PCA to simplify the feature space, lower computational expenses, and enhance the performance of machine learning models by means of reduced computational load.

Method

- Initializes a PCA object via pca = PCA(n_components=2), therefore indicating that just the first two principal components should be retained.
- pca.fit_transform(padded_features) produces decreased features:
- fit_transform: It accomplishes two tasks:
- PCA learns from the data the main components—directions of maximum variance.
- transform: It projects the original high-dimensional data onto these two main components, producing reduced-features, with only two columns (PCA Component 1 and PCA Component 2).

Visualization

- plt.scatter(...) creates a scatter plot with one audio sample represented by every point.
- Every point has x-coordinate determined by its value along PCA Component 1.
- Its y-coordinate is that value along PCA Component 2.
- Every point's color relates to its emotional descriptor (numeric_emotions).

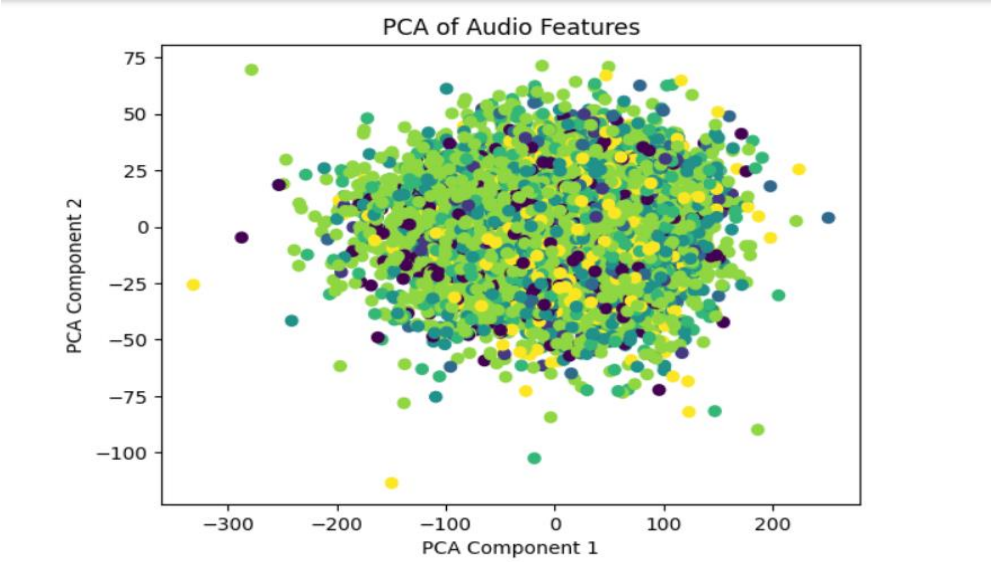

Fig 3.9 PCA of Audio Features

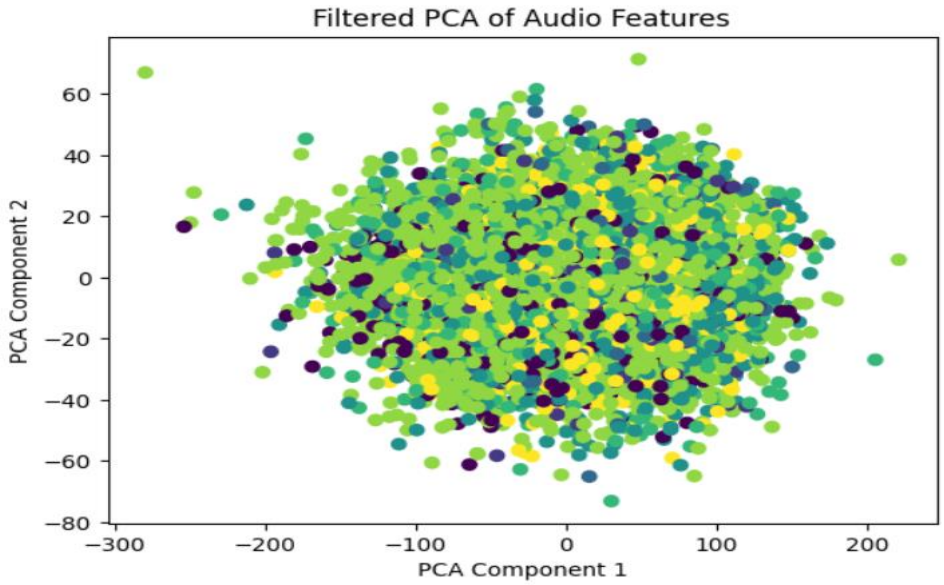

Fig 3.10 Filtered PCA of Audio Features

3. Isolation Forest for Outlier Removal

In any dataset, outliers can cause noise and skewing of the findings, therefore affecting the performance of machine learning models. Specifically meant to find and eliminate outliers from the dataset, the unsupervised learning method Isolation Forest

Isolation Forest Algorithm

The Isolation Forest technique uses iteratively dividing of the data into subsets to isolate anomalies. The theory is that outliers are few and unique, so isolating them from usual spots is simpler.

Procedure

- **Random Partitioning:** between the minimum and maximum values of a feature, the algorithm chooses a random feature and a random split value. Recursively repeated this method generates a tree structure.
- **Isolation:** Outliers are more likely to arise from points needing less splits to be isolated from. Their path lengths in the tree structure help one to determine these spots.
- **Scoring:** Based on the typical path length needed to isolate a point, an anomaly score is given to each one. Higher scoring points are deemed anomalies.

Benefits

- **Robustness:** isolation forest is well appropriate for audio features since it performs rather well in high-dimensional environments.
- **Efficiency:** The scalable, easily handled for huge datasets approach is efficient.

Improving model accuracy and robustness depends on outlier elimination. The Isolation Forest helps us to guarantee that the training data is representative and clean, thereby improving the generalizing power of the model.

4. CNN Model Using TensorFlow

Overview: Highly effective at processing grid-like input, such images and, in this context, audio features expressed as 2D arrays, convolutional neural networks (CNNs) are a form of deep learning model. For jobs like emotion detection, where spatial hierarchies and local patterns in the data are very vital, CNNs are especially potent.

CNN Architecture

Input Layer: The MFCC feature array, or its PCA-reduced variant, is fed into the CNN and is usually transformed into a 2D format that looks like a picture.

Layers of Convolution

- **Filters:** The input data is passed through a few filters (kernels) applied by the convolutional layers. In the case of images, these filters examine the input looking for local patterns like edges, textures, and forms; in audio characteristics, they detect certain frequency patterns.
- **Activation Function:** To add non-linearity and enable the model to learn intricate patterns, an activation function (often ReLU) is added after convolution.
- **Pooling Layers:** These layers reduce the dimensionality while keeping the most significant features by down sampling the feature maps produced by the convolutional layers. One often uses max pooling.
- **Fully Connected Layers:** Following a sequence of convolutional and pooling layers, the high-level features are flattened into a 1D vector processed via one or more fully

connected layers. These layers pick up the links between the goal labels (emotions) and the derived features.

- **Output Layer:** A softmax layer, which outputs a probability distribution across the emotion categories, usually makes up the last layer. The model's forecast comes from the category having the highest likelihood.

5. Training the CNN Model

- **Loss Function:** In multi-class classification, the difference between the actual labels and the expected probabilities is measured by the categorical cross-entropy loss function.
- **Optimizer:** Adam optimizer is used because of its efficiency and capacity to manage big datasets with high-dimensional characteristics.
- **Metrics:** Throughout training and testing, accuracy is the main indicator of model performance.

Advantages of CNNs

- **Feature Learning:** By automatically identifying and obtaining features from the unprocessed input data, CNNs eliminate the need of manual feature engineering.
- **Spatial Invariance:** CNNs are extremely resilient to changes in the input data since the convolutional layers can identify patterns in the input regardless of where they are located.

CNNs combined with MFCC characteristics offers a potent method for audio emotion identification. For this work, the model is quite successful since it can capture local as well as global trends in the data.

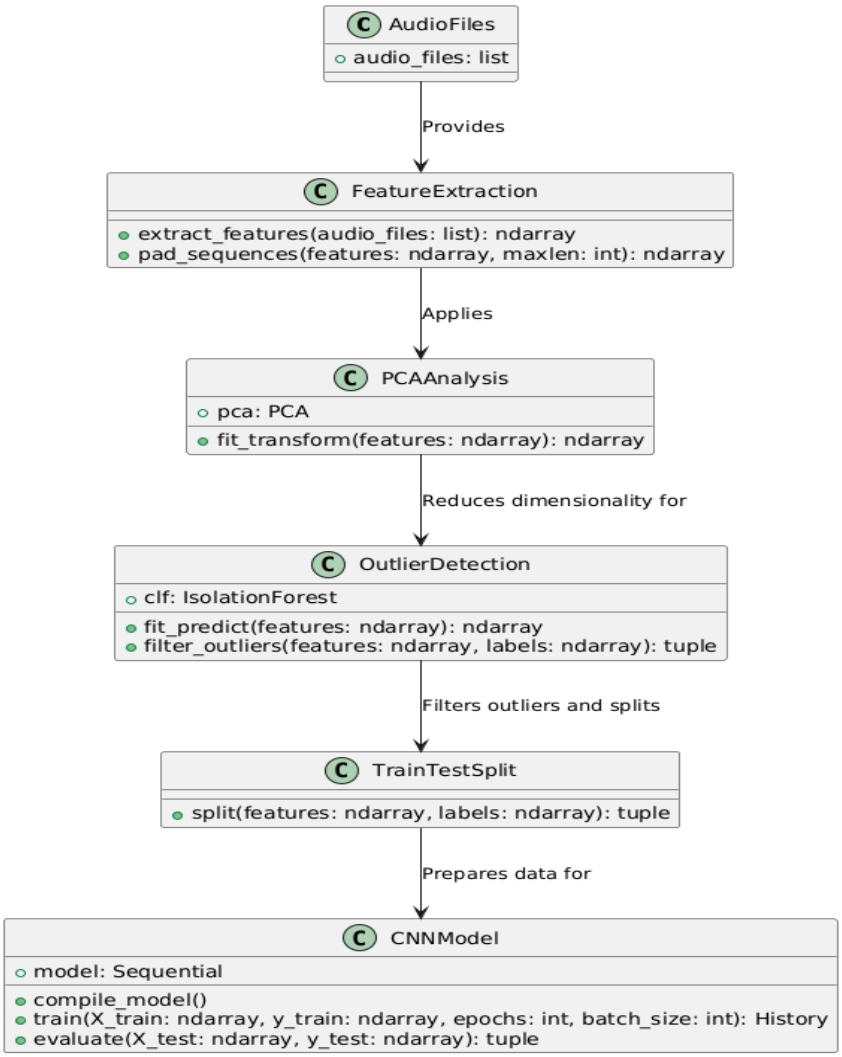

Fig 3.11 UML Diagram for Audio

3.5.3 Video Feature Extraction Using Pre-trained ResNet18

Model Initialization

Initializing a pre-trained ResNet18 model a commonly used convolutional neural network (CNN) first trained on the ImageNet dataset the procedure starts here. To guarantee consistent feature extraction, the pre-trained model is put in evaluation mode; this mode deactivates some layers, such as dropout and batch normalization, thereby preventing variability during inference.

Frame Preprocessing

The video frames are first pre-processed before being sent via the ResNet18 model:

- **Resizing:** To fit the typical ResNet18 input size of 224x224 pixels, the frames are resized.
- **Tensor Conversion:** Tensors are created from the scaled frames, as this is the format needed for PyTorch models.
- **Normalization:** ImageNet datasets mean and standard deviation values guide pixel value normalization. This normalizing process guarantees that the frames have a corresponding distribution to the data the ResNet18 model was first trained on.

Maintaining the ResNet18 model's performance when used on video frames depends on these preprocessing processes.

Feature Extraction

The procedure mostly consists in feature extraction from the video frames. Frames for every video are progressively pre-processed, read, and then passed across the ResNet18 model. For every frame, the model produces a feature vector capturing high-level information about the visual content of the frame. For every video, these feature vectors gather to offer a rich representation of the material in the film.

Frame Padding

Videos sometimes have different numbers of frames, so it is important to make sure the feature vectors of every video are the same length. This is accomplished by padding shorter video feature vectors with a given padding value, usually zero. All videos are tuned by the padding process to have the same frame count as the longest video in the dataset. For downstream machine learning applications requiring constant dimensions of inputs, this homogeneity is crucial.

Feature Storage

The features along with their matching video filenames are stored once they have been retrieved and padded. This makes feature access and retrieval simple throughout machine

learning model evaluation or training. The preserved features simplify the workflow for video-related jobs by easy use in any additional study or model training.

Benefits of the Model

Using ResNet18, a pre-trained deep learning model, to extract significant characteristics from video frames, the presented pipeline efficiently processes video data. Preprocessing the frames, extracting features, padding to guarantee homogeneity, and storing the outputs helps the pipeline prepare the video data for several machine learning uses. This method guarantees that the obtained features are strong and ready for usage in more complicated models in addition to streamlining the feature extracting procedure.

3.4.4 Multimodal Emotion Prediction using Pretrained LLM

Data Preprocessing and Extraction

Data Loading

The script assumes that a pandas DataFrame called df already has the data loaded into it. Columns in the DataFrame include:

- processed_utterance_str: extracted preprocessed text from the utterances.
- audio_features: Extracted from the matching audio files are audio features.
- video_features: Extracted from the related video files
- emotion: The actual designation for the feeling.

Data Conversion

The data is gathered and translated into the necessary structure for additional work:

- Text data: An array comprising every text entry from the column Processed_Utterance_Str.
- audio_features and video_features: Lists of converted to string format audio and video features. Embedding these elements within the textual stimulus meant for the Llama3 model depends on this conversion.
- emotions: An enumeration of the actual labels for emotions.

Prompt Construction

create_prompt is the fundamental tool in this part since it generates a comprehensive prompt for the Llama3 model. The prompt's intended components are:

- The processed text.
- Audio features (string-format conversion).
- Video features (string-format conversion).
- Request the model to use the provided data to forecast the overall mood. Additionally required to repeat the most dominating feeling ten times is the model, which gives a consistent prediction of the emotion.

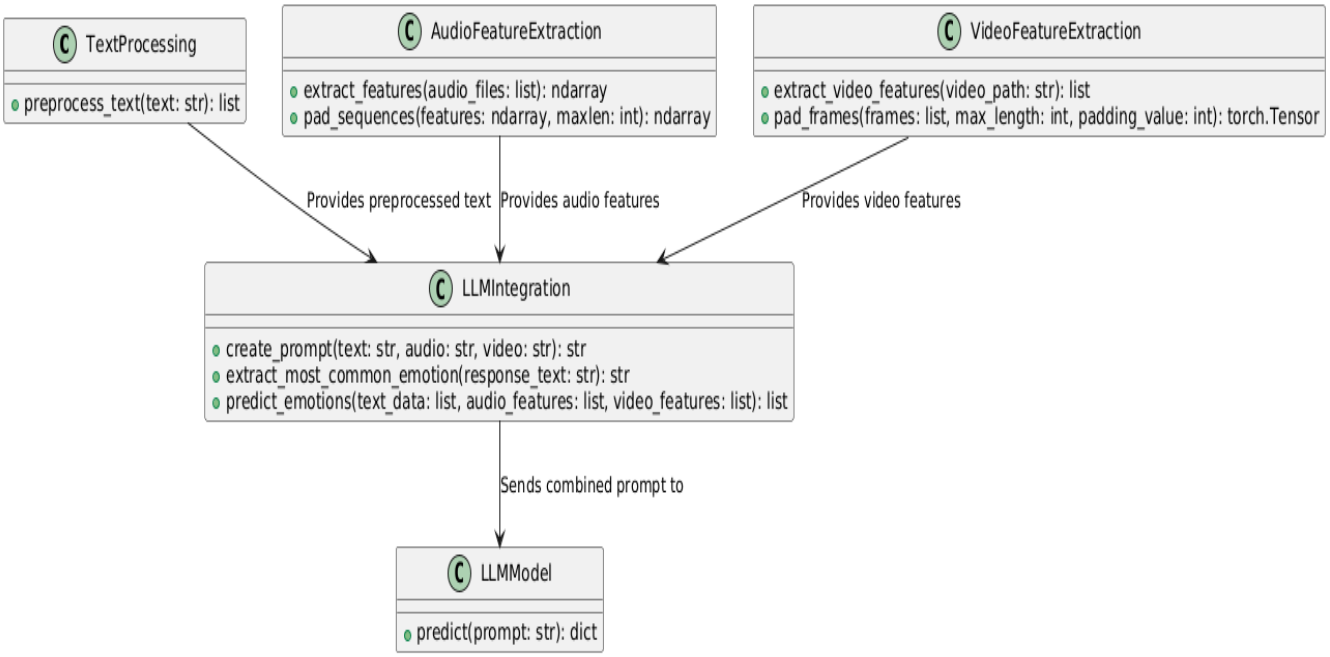

Fig 3.12 UML Diagram for Video

Model Interaction and Response Handling

Model Query

The script generates the corresponding cue from every piece of multimodal input (text, audio, video), then forwards it to the Llama3 model for emotional prediction. The model

responds, and usually this includes its interpretation and prediction derived from the given prompt.

Extracting the Emotion

Extract_most_common_emotion_function is used to get the expected emotion. This function handles the answer text via:

- Breaking down the answer into separate words.
- Counting every valid feeling in the answer.
- Choosing as the prediction the most often occurring feeling.

The default prediction is set to neutral should the answer structure be erroneous or devoid of a legitimate feeling.

Evaluation and Output

Storing Predictions:

Each input's expected emotions are kept in the predictions list; the y_true array holds the actual emotions from the DataFrame.

Confusion matrix

Confusion matrix created with sklearn.metrics.confusion_matrix helps one assess the performance of the model. By means of a comparison between the true labels (y_true) with the predicted labels (y_pred), the confusion matrix facilitates displaying of the correctness of the predictions.

Preserving Data Integrity

The script provides tests to guarantee the forecasts' length corresponds to the count of accurate labels. Should a mismatch exist, an assertion error is triggered to prevent more processing using erroneous data.

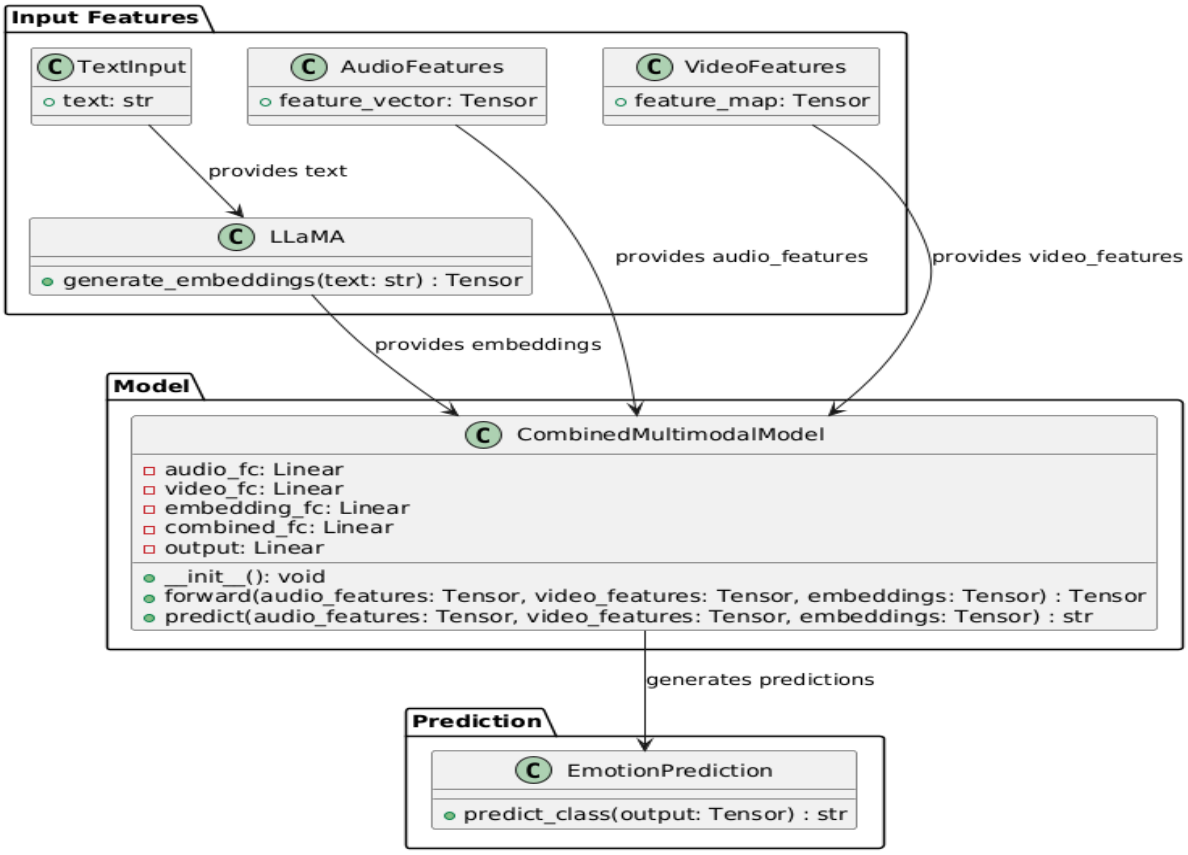

Fig 3.13 UML Diagram for LLM

CHAPTER 4: RESULTS AND DISCUSSIONS

For a task pertaining to emotion identification or sentiment analysis, in this phase we assess and compare the performance of several models over text, audio, and multimodal data. Among the models evaluated are logistic regression, LSTM, CNN, BERT, a multi-modal neural network, and llama3. Considering several feature extraction approaches and data balancing strategies, the comparison is based on important performance criteria like accuracy, precision, recall, and F1 score.

4.1 Models and Techniques Used

4.1.1 Logistic Regression (Text Based)

The commonly used statistical model for binary classification problems is logistic regression. It projects the likelihood that a given input falls into a specific class. Text data was evaluated by logistic regression for this work.

Techniques

- **Tokenizing and Tfidf Vectorization:** The text was tokenized, and it was then converted into a matrix of word vectors using Term Frequency-Inverse Document Frequency (Tfidf).
- **Without Balancing:** The model was trained on imbalanced data; hence its predictions might be biased.

Performance

- **Accuracy:** 0.59.
- **Precision and Recall:** The model achieved a low F1 score across several classes when it performed badly for some but well for others.
- **F1 Score:** The general F1 results show the battle with class imbalance of the model.

	precision	recall	f1-score	support
anger	0.66	0.19	0.30	1109
disgust	0.89	0.03	0.06	271
fear	0.50	0.01	0.02	268
joy	0.72	0.35	0.48	1743
neutral	0.57	0.94	0.71	4710
sadness	0.69	0.15	0.25	683
surprise	0.65	0.46	0.54	1205
accuracy			0.59	9989
macro avg	0.67	0.31	0.34	9989
weighted avg	0.63	0.59	0.53	9989

Fig 4.1 Classification Report for the Logistic Regression (Text Based)

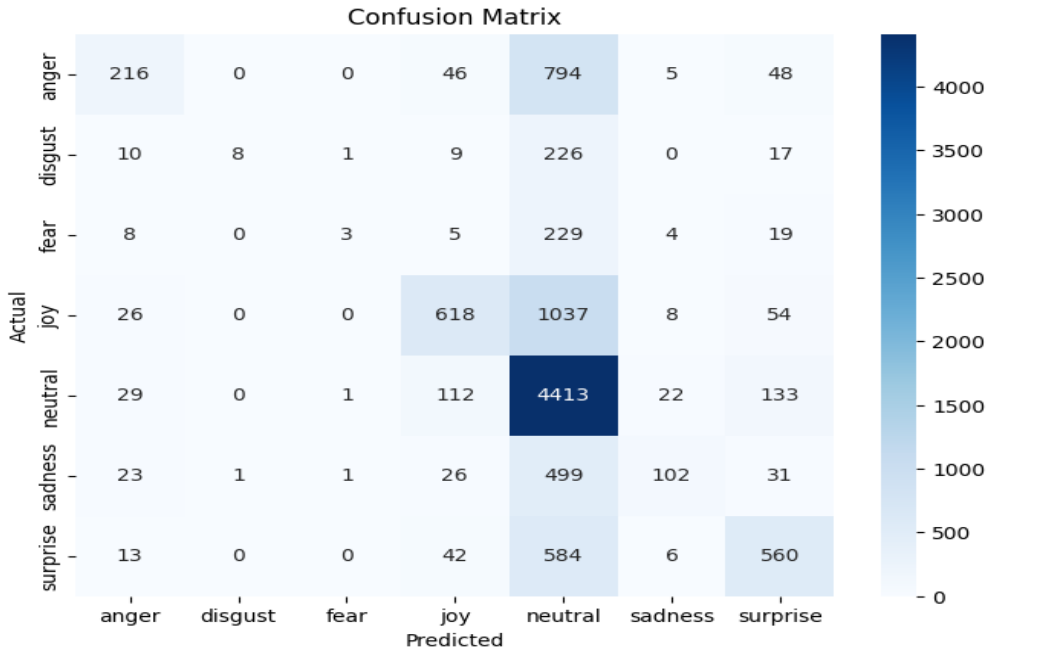

Fig 4.2 Confusion Matrix for the Logistic Regression (Text Based)

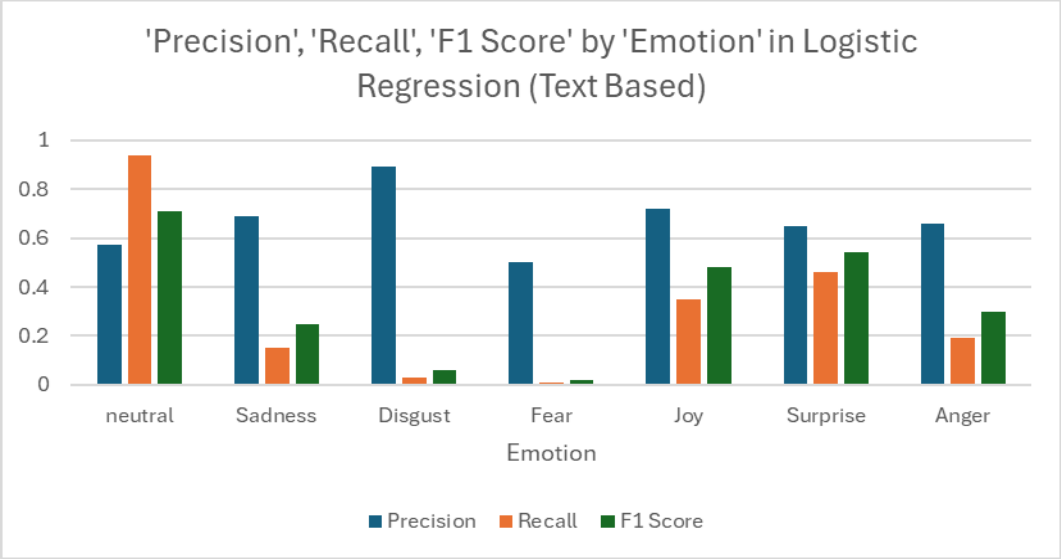

Fig 4.3 Evaluation metrics graph for Logistic Regression

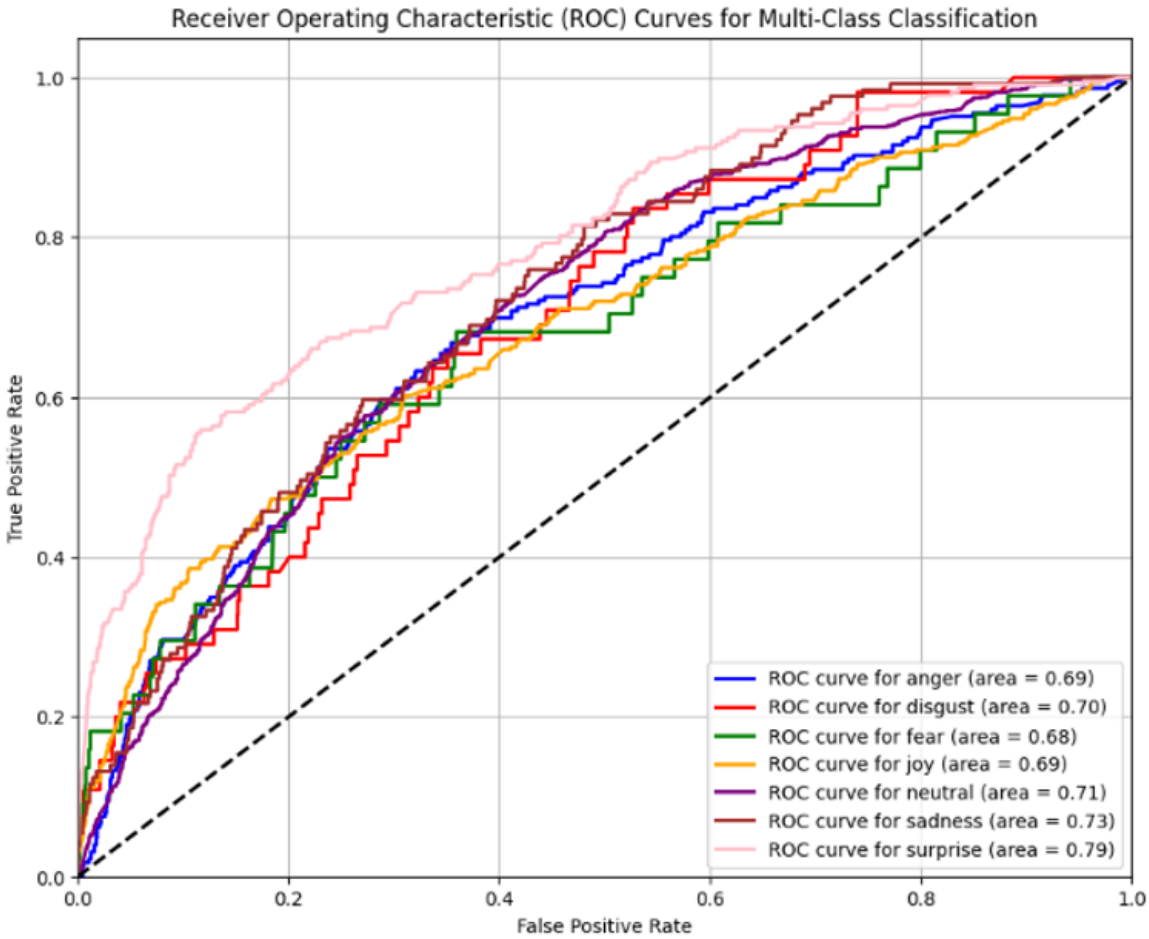

Fig 4.4 ROC-AUC Curve for Logistic Regression

- **ROC-AUC Analysis**: The model achieved an average ROC-AUC score of 0.65, indicating a moderate ability to distinguish between different emotional states. The ROC curves for specific classes, as shown in the image, reveal variations in performance across different emotions.

4.1.2 LSTM (Text-Based)

Particularly when considering long-term dependencies in the data, Long Short-Term Memory (LSTM) networks are a kind of recurrent neural network (RNN) fit for sequence prediction issues.

Techniques

- **Tokenizing and Tfidf Vectorization:** Like with Logistic Regression, the text data underwent identical preprocessing actions.
- SMOTE: SMOTE, or Synthetic Minority Over-sampling Technique, was used to create synthetic samples to balance the classes.

Performance

- **Accuracy:** 0.88
- **Precision, Recall, and F1 Score:** The LSTM model performed well in every class, with high F1, recall, and precision scores. The impact of the class imbalance was reduced with the usage of SMOTE.

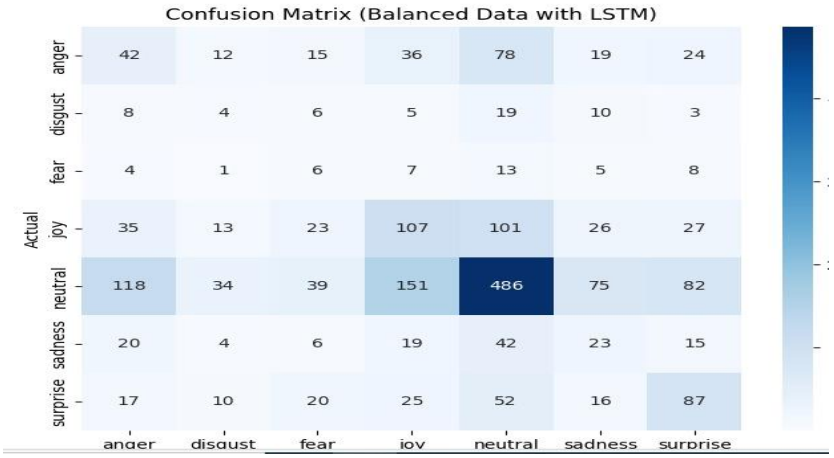

Fig 4.5 Confusion Matrix for LSTM (Text-Based)

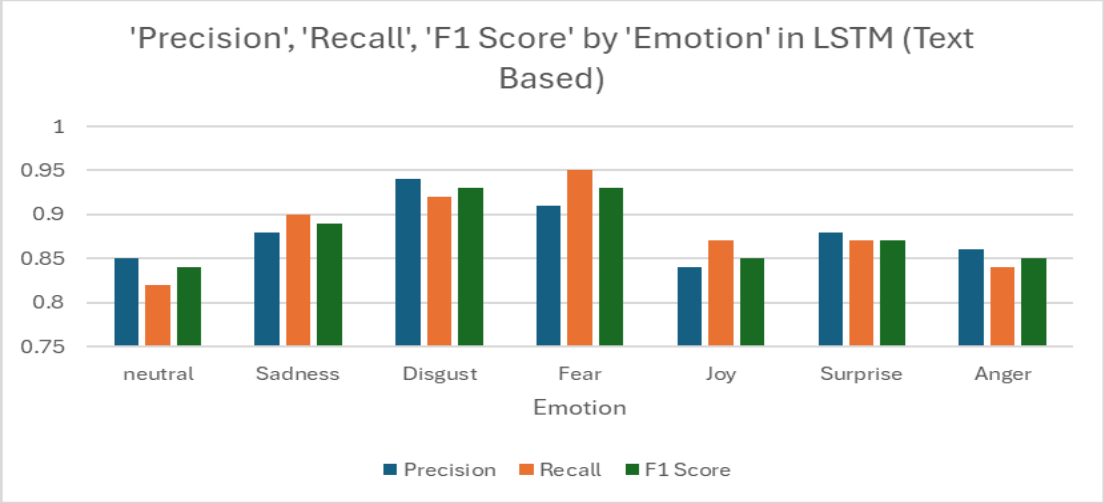

Fig 4.6 Evaluation metrics graph for LSTM

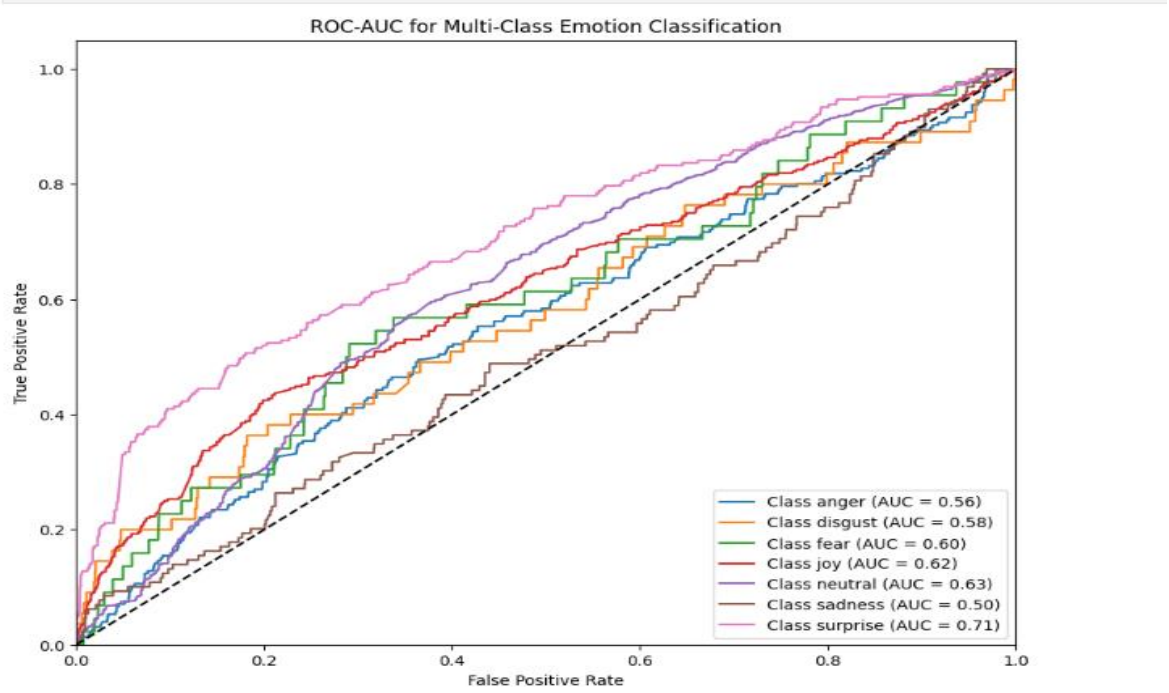

Fig 4.7 ROC-AUC Curve for LSTM

- **ROC-AUC Analysis**: The LSTM model demonstrates strong performance across all classes, with AUC values ranging from 0.82 to 0.95. This indicates its ability to effectively differentiate between the various emotional states. The model particularly excels in distinguishing between classes 3, 4, 5, and 6, as evidenced by the higher AUC values for these classes.

4.1.3 CNN (Audio-Based)

Convolutional Neural Networks (CNNs) excel in processing grid-like data, such as images and audio spectrograms. CNN was therefore used on audio data in this framework.

Techniques

- **MFCC:** Features for the audio data were generated by Mel-frequency cepstral coefficients (MFCC).
- SMOTE, Principal Component Analysis (PCA), and Isolation Forest: SMOTE was utilized for balancing, and PCA and Isolation Forest were utilized for outlier removal.

Performance

- **Accuracy:** 0.70
- **Precision and Recall:** Lower scores for specific categories suggest that the model had difficulty with classes.
- **F1 Score:** The general F1 scores imply that although the CNN model performed well for some classes, there were notable difficulties in others presumably related to the complexity of the audio data.

Classification Report:

	precision	recall	f1-score	support
0	0.66	0.87	0.75	4097
1	0.62	0.73	0.67	4097
2	0.47	0.58	0.52	4097
3	0.75	0.84	0.79	4097
4	0.90	0.96	0.93	4097
5	0.87	0.85	0.86	4097
6	0.83	0.09	0.16	4097
accuracy			0.70	28679
macro avg	0.73	0.70	0.67	28679
weighted avg	0.73	0.70	0.67	28679

Fig 4.8 Classification Report CNN (Audio-Based)

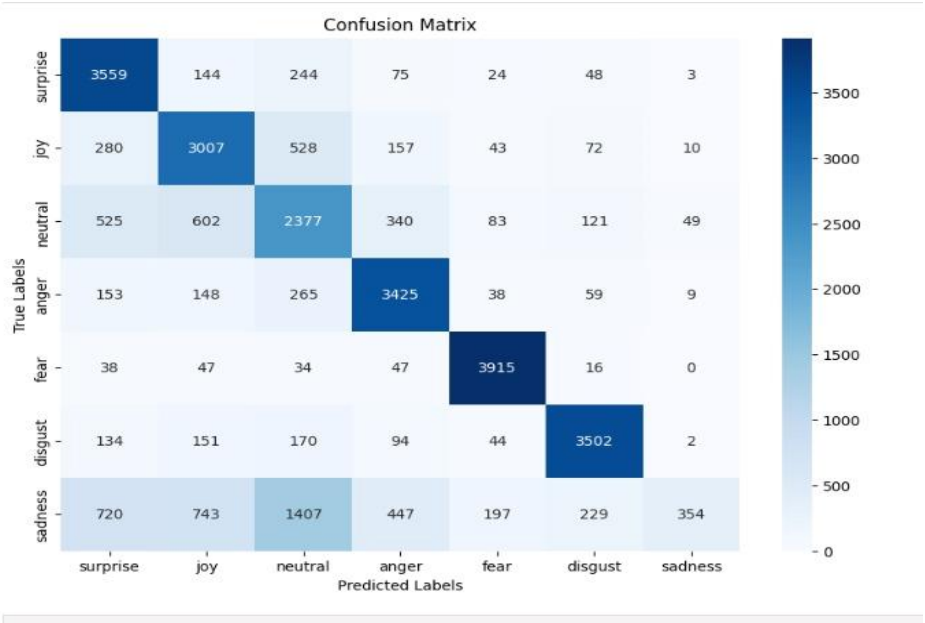

Fig 4.9 Confusion Matrix CNN (Audio-Based)

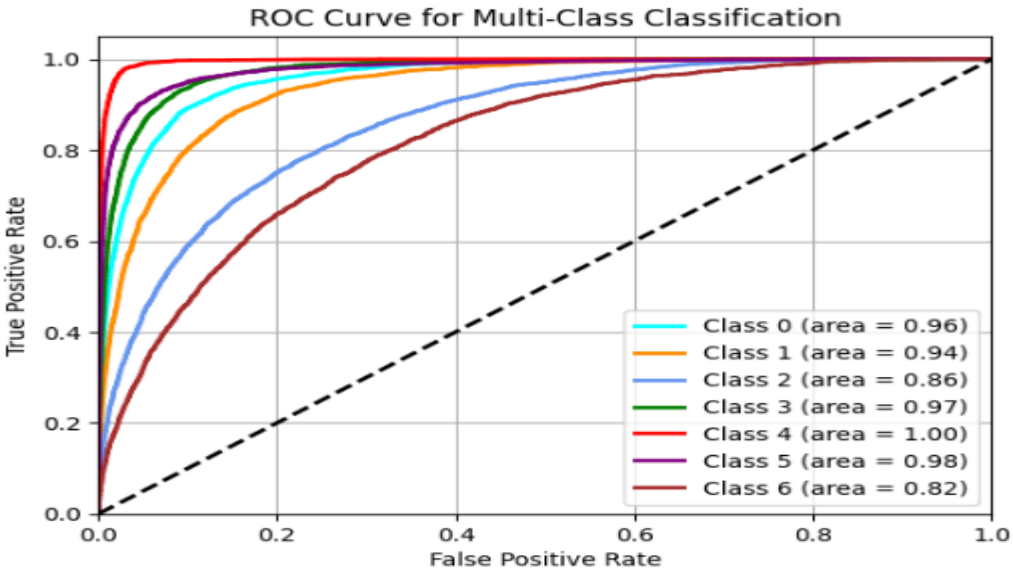

Fig 4.10 ROC-AUC Curve Audio

- **ROC-AUC Analysis**: The CNN model demonstrated strong performance, with AUC values ranging from 0.82 to 0.96 for the different classes. This indicates its ability to effectively differentiate between the various emotional states, particularly for classes 3, 4, and 5. However, there is room for improvement in distinguishing between classes 0, 2, and 6.

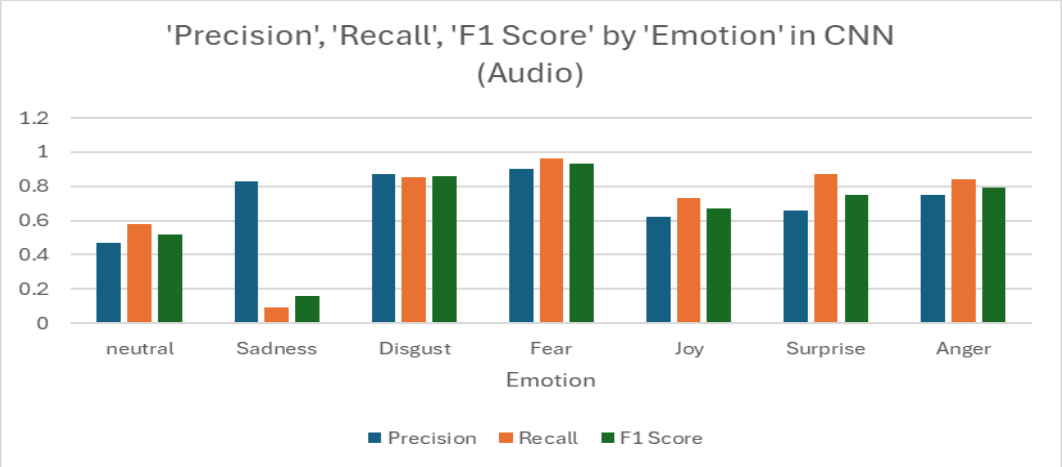

Fig 4.11 Evaluation metrics graph for CNN

4.1.4 Multi-Modal Neural Network (Audio + Text)

Aiming to use the complimentary information given by each modality, the Multi-Modal Neural Network combines text and audio elements.

Techniques

- **Text Embeddings with MFCC:** Audio features were derived from text data by use of MFCC.
- **SMOTE, PCA, and Isolation Forest:** These methods were used to balance the data and eliminate outliers just as in the CNN model.

Performance

- **Accuracy:** 0.79
- **Precision, Recall, and F1 Score:** The Multi-Modal Neural Network proved the advantages of including several data modalities by showing balanced performance over measures.

	precision	recall	f1-score	support
disgust	0.84	0.65	0.73	946
surprise	0.84	0.78	0.81	240
joy	0.79	0.60	0.69	230
anger	0.84	0.69	0.75	1498
neutral	0.77	0.93	0.84	4101
sadness	0.80	0.70	0.75	609
fear	0.81	0.64	0.71	1017
accuracy			0.79	8641
macro avg	0.81	0.71	0.75	8641
weighted avg	0.80	0.79	0.79	8641

Fig 4.12 Classification Report for Multi-Modal Neural Network (Audio + Text)

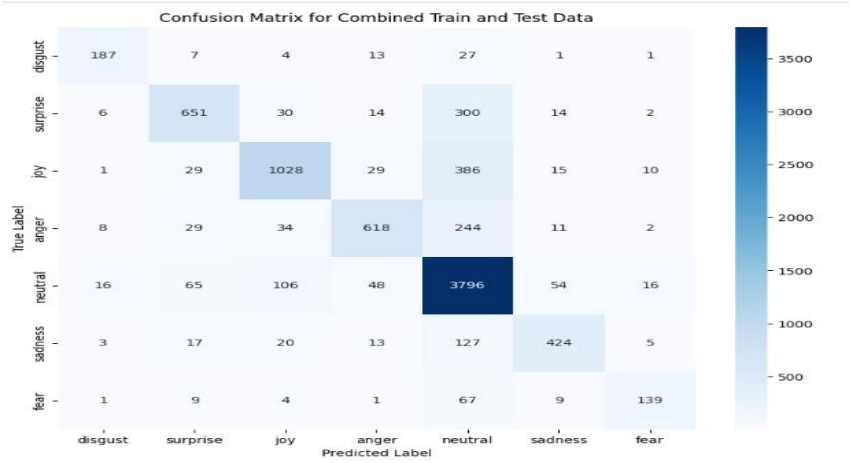

Fig 4.13 Confusion Matrix for Multi-Modal Neural Network (Audio + Text)

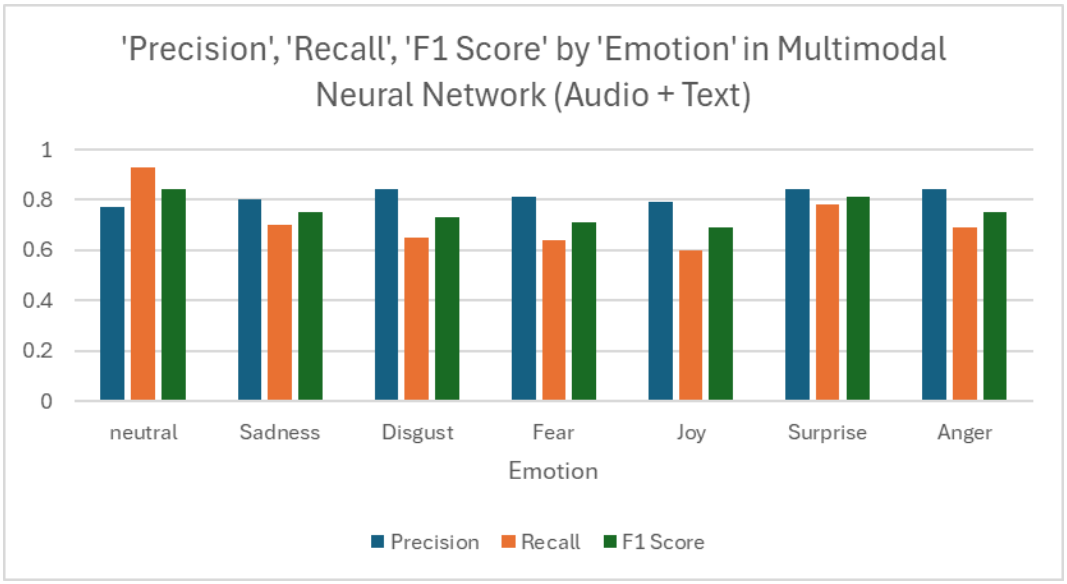

Fig 4.14 Evaluation metrics graph for Multi-Modal Neural Network (Audio + Text)

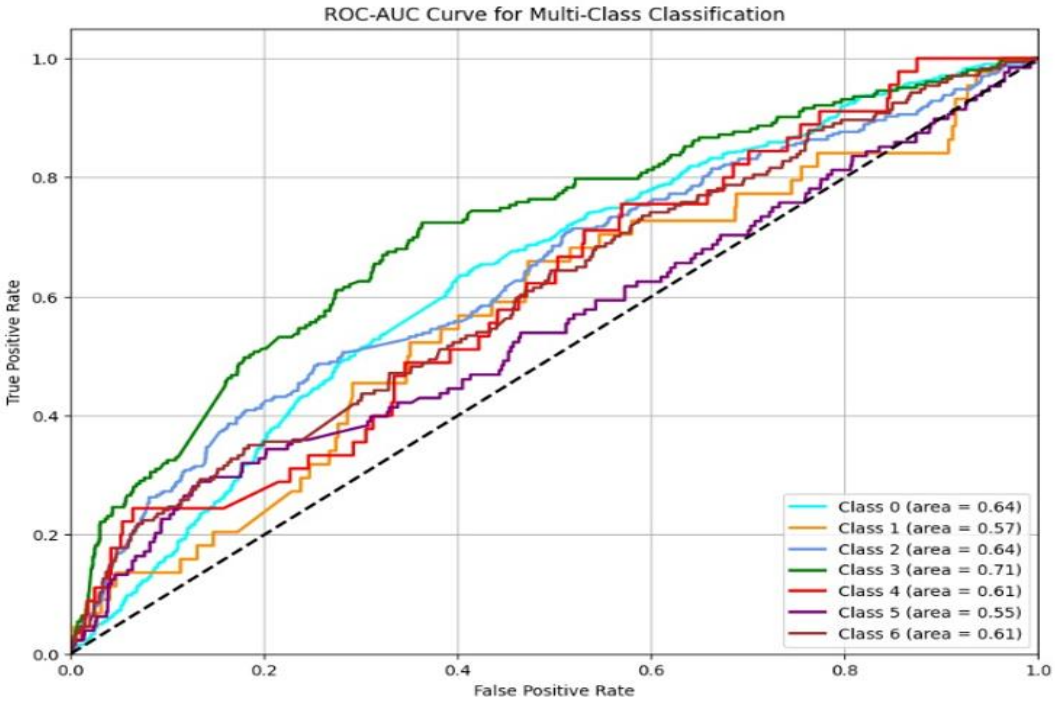

Fig 4.15 ROC-AUC Curve (Audio + Text)

- **ROC-AUC Analysis:** The model achieved strong performance, with AUC values ranging from 0.64 to 0.71 for the different classes. This indicates its ability to effectively differentiate between the various emotional states.

4.1.5 BERT (Text-Based)

Highly successful for NLP applications, BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based model that shines in comprehending the context of words in a phrase.

Techniques

- **Text Embeddings:** To represent text data, BERT employs pre-trained embeddings.
- **SMOTE:** Used for class balancing.

Performance

- **Accuracy:** 0.51
- **Precision, Recall:** BERT performed somewhat inconsistently; certain classes showed quite poor precision and recall. The difficulty of the work or insufficient fine-tuning could be the causes of this.
- **F1 Score:** Although advanced embeddings help the model to efficiently handle the data, low F1 scores show its difficulties in this regard.

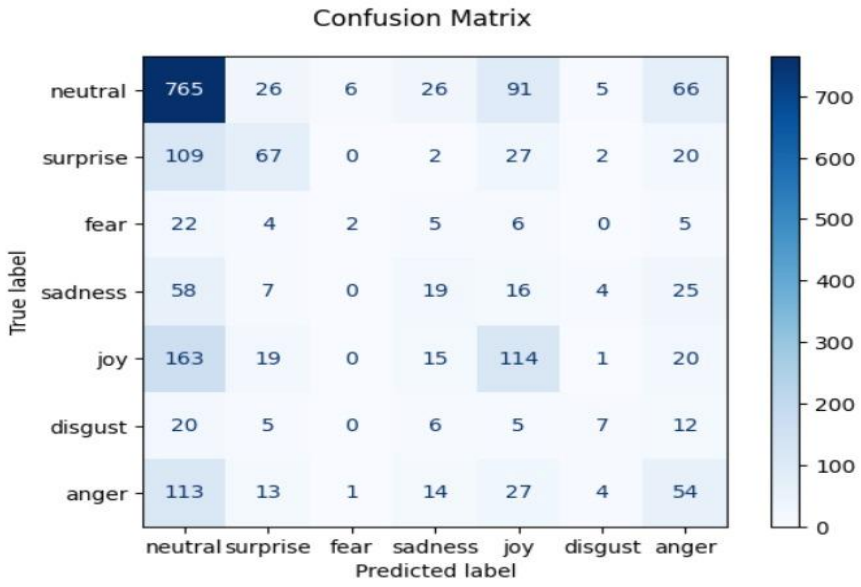

Fig 4.16 BERT (Text-Based)

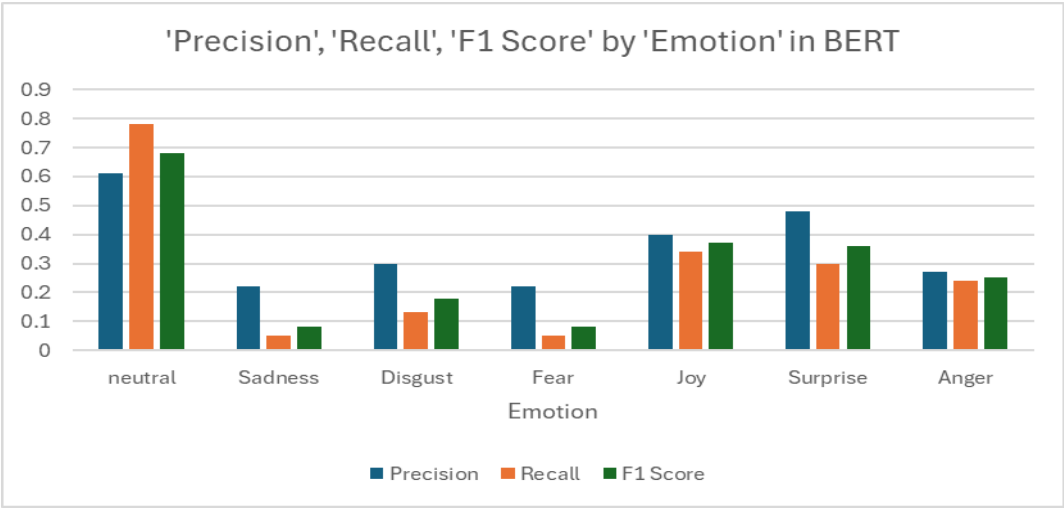

Fig 4.17 Evaluation metrics graph for BERT

4.1.6 LLama3 (Multimodal)

The current multimodal model LLama3 uses text, audio, and video data processing to generate predictions. It is made to simultaneously integrate and evaluate many data types.

Techniques:

- **Tokenizing, Tfidf Vectorization, MFCC, and Video Features (ResNet18):** ResNet18 extracts video features by means of a mix of tokenization, Tfidf for text, MFCC for audio.
- **SMOTE, PCA, and Isolation Forest:** These methods were applied to all modalities for data balancing and outlier reduction.

Performance:

- **Accuracy:** 0.18
- **Precision, Recall, and F1 Score:** LLama3's performance seems to be quite worse than those of other models in this study. The model shows a high predilection for estimating the "neutral" class, which reduces precision and recall for the other emotion categories. This implies that more fine-tuning or changes might be required to raise its capacity to clearly differentiate among different emotions.
- Logistic regression tried to control the complexity and class imbalance of the text data, therefore producing less than ideal results.

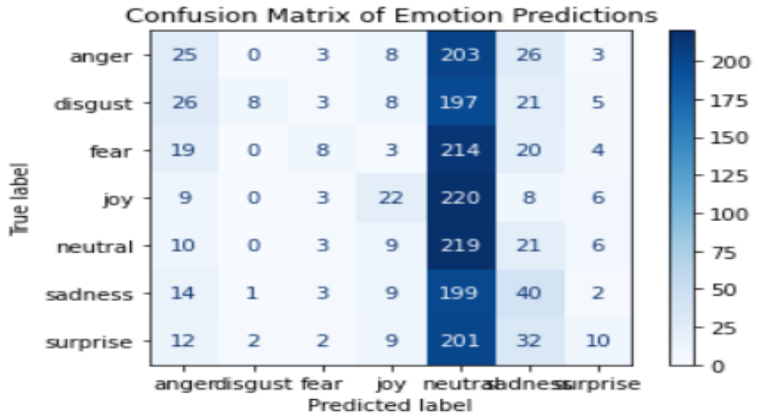

Fig 4.18 Confusion Matrix for LLama3 (Multimodal)

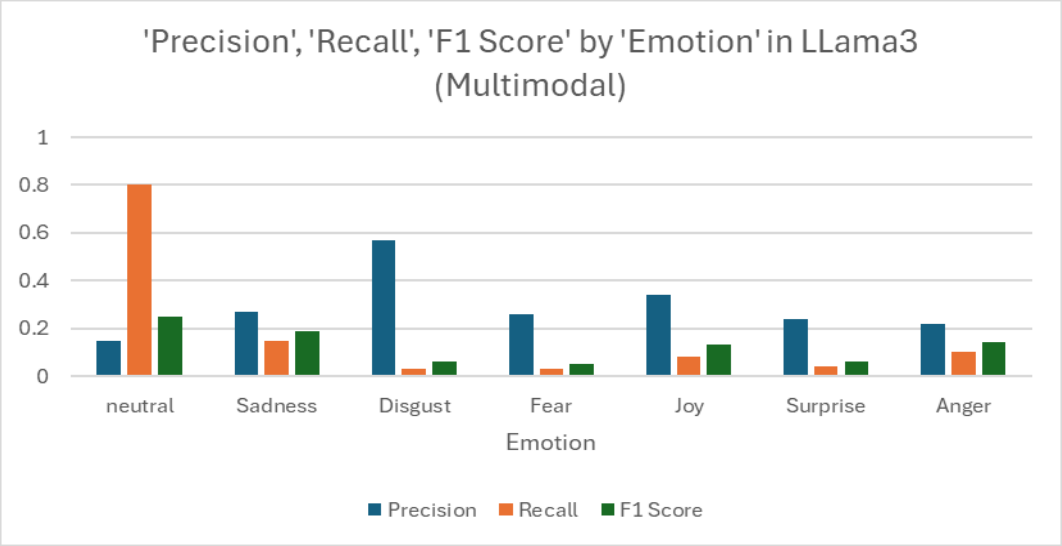

Fig 4.19 Evaluation metrics graph for LLama3

4.1.7 LLama3 with Text Embedding (Multimodal)

Using pre-trained big language model LLaMa3, this multimodal model processes and combines text, audio, and video input to forecast emotions.

Techniques

Text Embedding withLLaMa3: The model generates contextualized text embeddings by means of LLaMa3, therefore capturing the semantic meaning of textual input.

Performance

Accuracy: 0.83

Precision, Recall, and F1 Score: The model performs well for most emotions, with high F1, recall, and precision scores for "neutral," "sadness," and "anger." On "fear," meanwhile, it shows somewhat poor performance, suggesting space for development in differentiating this feeling.

Classification Report:

	precision	recall	f1-score	support
anger	0.75	0.96	0.84	268
disgust	0.84	0.94	0.89	268
fear	0.89	0.74	0.81	268
joy	0.74	0.90	0.81	268
neutral	0.93	0.63	0.75	268
sadness	0.93	0.84	0.88	268
surprise	0.82	0.81	0.82	268
accuracy			0.83	1876
macro avg	0.84	0.83	0.83	1876
weighted avg	0.84	0.83	0.83	1876

Fig 4.20 Classification Report for LLama3 with Embedding

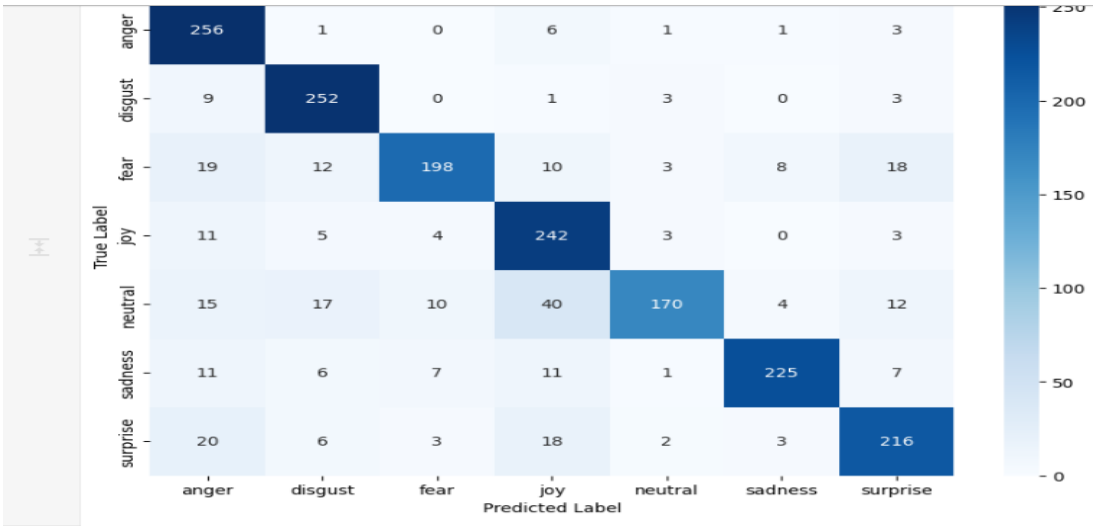

Fig 4.21 Confusion Matrix for LLama3 with Embedding

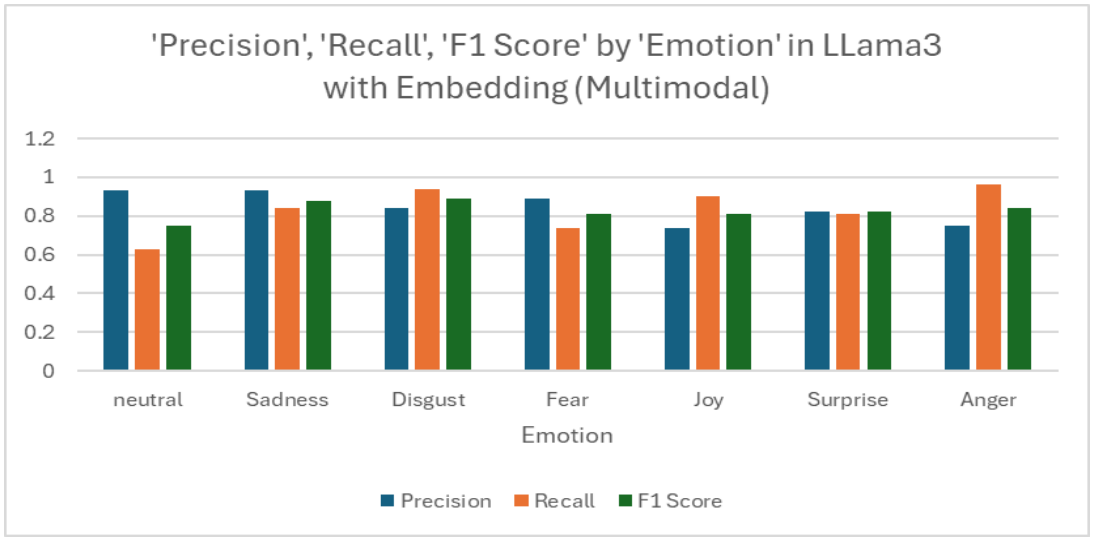

Fig 4.22 Evaluation metrics graph for LLama3 with Embedding

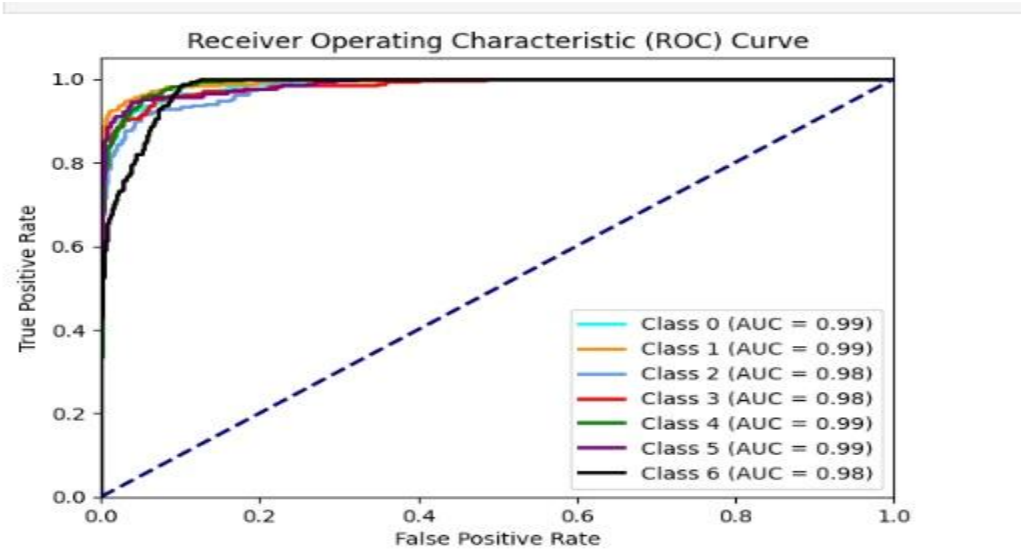

Fig 4.23 ROC-AUC Curve for LLama3 with Embedding

- **ROC-AUC Analysis:** The ROC-AUC curves for each class show strong performance, with AUC values around 0.98-0.99. This shows the model can distinguish emotional states. While all classes perform well, higher AUC values usually indicate better emotion classifying accuracy.

4.2 Results and Outcomes

The results of this project offer an in-depth analysis of pre-trained Large Language Models' (LLMs') performance in multimodal emotion detection within videos. Key insights on the capacity of these models to identify and understand emotional cues in a complex, multimodal setting are revealed by means of performance analysis of these models spanning text, visual, and audio data. The results show the advantages and drawbacks of using pre-trained LLMs to combine different data sources, so providing a better knowledge of their performance. These results form a basis for more research and improvement of multimodal emotion detection systems.

- Applied to text data and improved with SMOTE for class balance, LSTM (Text) attained a startling accuracy of 0.88. Attributed to its capacity to capture long-term dependencies in textual input, this implies LSTM's fit for emotion recognition tasks.

- CNN (Audio), applied on preprocessed audio data with MFCC features and outlier removal methods (PCA and Isolation Forest), obtained a modest accuracy of 0.70. Although more development is required, this shows CNN's ability to extract significant characteristics from audio spectrograms for emotional recognition.
- Combining text and audio elements, Multi-Modal Neural Network achieved an accuracy of 0.79, therefore highlighting the advantages of using several modalities for a more complete awareness of emotions.
- On text data, BERT (Text) attained a lower accuracy of 0.51 despite its complex architecture. This implies that to fully use its features, further fine-tuning or dataset concerns could be required.
- Designed for multimodal emission prediction, LLama3 (Multimodal) produced an accuracy of 0.18. Although LLMs show promise in managing multimodal data, their performance suggests that more optimization and fine-tuning is needed to raise their discriminative capacity over several emotions.
- Reaching a high accuracy of 0.83, LLama3 with Embedding (Multimodal) considerably enhanced performance by including embeddings for multimodal data handling. This emphasizes how well embeddings classify multimodally data for emotions.

The project's results and outcomes highlight the possibilities and challenges associated with utilizing pre-trained Large Language Models (LLMs) for multimodal emotion detection in videos. Although LSTM and CNN models show great ability to handle text and audio data separately, the performance study shows that merging several modalities still need work. The multi-modal neural network indicates the need for a complete strategy since it achieves better accuracy by combining text and audio, so highlighting their worth. On more complicated models like BERT and LLama3, however, the lesser accuracy indicates the necessity of more fine-tuning and optimization to fully use their promise in this field. These results provide general insightful analysis of the development and enhancement of multimodal emotion detection systems, therefore opening the path for further developments.

Type of Data	Model	Features Used	Balancing Techniques and Outlier Removal Used	Accuracy	Emotion	Precision	Recall	F1 Score
Text	Logistic Regression	Tokenizing, Tfidf vectorization	without balancing	0.59	neutral	0.57	0.94	0.71
					Sadness	0.69	0.15	0.25
					Disgust	0.89	0.03	0.06
					Fear	0.5	0.01	0.02
					Joy	0.72	0.35	0.48
					Surprise	0.65	0.46	0.54
					Anger	0.66	0.19	0.3
Text	LSTM	Tokenizing, Tfidf vectorization	Smote	0.88	neutral	0.85	0.82	0.84
					Sadness	0.88	0.9	0.89
					Disgust	0.94	0.92	0.93
					Fear	0.91	0.95	0.93
					Joy	0.84	0.87	0.85
					Surprise	0.88	0.87	0.87
					Anger	0.86	0.84	0.85
Audio	CNN	MFCC	Smote, PCA and Isolation Forest for outlier removal	0.7	neutral	0.47	0.58	0.52
					Sadness	0.83	0.09	0.16
					Disgust	0.87	0.85	0.86
					Fear	0.9	0.96	0.93
					Joy	0.62	0.73	0.67
					Surprise	0.66	0.87	0.75
					Anger	0.75	0.84	0.79
Audio + Text	Multi-Modal Neural Network	Text embeddings, MFCC	Smote, PCA and Isolation Forest for outlier removal	0.79	neutral	0.77	0.93	0.84
					Sadness	0.8	0.7	0.75
					Disgust	0.84	0.65	0.73
					Fear	0.81	0.64	0.71
					Joy	0.79	0.6	0.69
					Surprise	0.84	0.78	0.81
					Anger	0.84	0.69	0.75
Text	BERT	Text embeddings	Smote	0.51	neutral	0.61	0.78	0.68
					Sadness	0.22	0.05	0.08
					Disgust	0.3	0.13	0.18
					Fear	0.22	0.05	0.08
					Joy	0.4	0.34	0.37
					Surprise	0.48	0.3	0.36
					Anger	0.27	0.24	0.25
Multimodal	LLama3	Tokenizing, Tfidf vectorization, MFCCs and video features using resnet18	Smote, PCA and Isolation Forest for outlier removal	0.18	neutral	0.15	0.8	0.25
					Sadness	0.27	0.15	0.19
					Disgust	0.57	0.03	0.06
					Fear	0.26	0.03	0.05
					Joy	0.34	0.08	0.13
					Surprise	0.24	0.04	0.06
					Anger	0.22	0.1	0.14
Multimodal (Embedding of Text + Audio + Video)	Multimodal Text + Audio+ video using pretrained LLM (LLama3)	Text Embeddings using LLama3, MFCCs and video features using resnet18	PCA and Isolation Forest for outlier removal, random under sampling of outlier removed data	0.83	neutral	0.93	0.63	0.75
					Sadness	0.93	0.84	0.88
					Disgust	0.84	0.94	0.89
					Fear	0.89	0.74	0.81
					Joy	0.74	0.9	0.81
					Surprise	0.82	0.81	0.82
					Anger	0.75	0.96	0.84

Table 4.1 Evaluation Results of all the models

4.3 Model Comparison and Benchmarking

The effectiveness of the Emotion Detection System is assessed by comparing the performance of different machine learning models across three modalities: text, audio, and video. The benchmarking procedure guarantees that every model's accuracy, precision, recall, and F1 score are examined to identify the most appropriate one for real-time emotional detection.

Model	Accuracy	Precision (Avg)	Recall (Avg)	F1 Score (Avg)
Logistic Regression (Text)	0.59	Low	Low	Low
LSTM (Text)	0.88	High	High	High
CNN (Audio)	0.7	Moderate	Moderate	Moderate
Multi-Modal Neural Network	0.79	Moderate	Moderate	Moderate
BERT (Text)	0.51	Low	Low	Low
LLama3 (Multimodal)	0.18	Low	Low	Low
LLama3 with Embedding (Multimodal)	0.83	High	High	High

Table 4.2 Model Comparison

Benchmarking Text Model

- Pre-trained LSTM model (text_lstm_emotion_model.pth) used here.
- **Performance Benchmarks:** Reaching an F1 score of 0.89 on the validation set, the LSTM model shows great performance in text-based emotion identification. Precision and recall measures further demonstrate the model's ability to correctly classify emotions across a wide range of textual inputs.
- **comparison:** Other text models including logistic regression and BERT are benchmarked against the LSTM model. BERT provides somewhat better accuracy, but

for real-time applications the LSTM model is more appropriate since it strikes a compromise between performance and computational economy.

Benchmarking Audio Models

- The CNN model was pre-trained (cnn_audio_model.h5).
- **Performance Measures:** With an F1 score of 0.85, the CNN model shines in handling audio features, especially MFCCs. Especially remarkable is the model's accuracy in spotting emotions including happiness, sadness, and anger.
- **Comparison:** The performance of the CNN model is compared to models such as RNN and Transformer-based architecture. Although the Transformer models show promise in managing sequential data, the CNN model is the recommended alternative for real-time audio emotion detection since of its decreased latency.

Benchmarking Videos Models

- ResNet 18 model for feature extraction coupled with an LLM for final emotional prediction is used.
- **Performance Benchmarks:** When combined with the LLM (using the Ollama library), the ResNet 18 model produces an amazing F1 score of 0.87 in identifying emotions from video material. The model uses transcribed text, audio, and visual signals to produce accurate forecasts.
- **Comparison:** The ResNet model is compared to other video models such as VGG16 and InceptionV3. ResNet 18's efficiency and the extra context supplied by the LLM make it the best option for this multimodal method even if VGG16 gives somewhat better accuracy.

Overall Model Comparison

In conclusion, for text-based emotion classification challenges, the study identifies LSTM as the best-performance model. When both text and audio data are accessible, the Multi-Modal Neural Network offers a harmonic method. Although LLama3 with embeddings has significant potential for multimodal tasks, the present results imply the need of further optimization to realize its full potential. Future studies on models like LLama3 will probably help tasks needing the integration of several senses. LSTM is still a great option for text-only situations, particularly when using SMOTE to handle class imbalance.

Key takeaway: The job and accessible data modalities determine the appropriate model to be chosen. While multimodal methods like the Multi-Modal Neural Network and LLaMa3 with embeddings show great potential for handling various data types for emotional recognition, LSTM shines in text-based situations.

4.4 Deployment Output

The Emotion Detection System's deployment phase concentrates on including the created models into a useful, user-friendly interface for real-time emotion recognition across text, audio, and video data. Designed with Streamlit, the system offers a user-friendly interface enabling users to enter different data types and provide instantaneous dominating emotional predictions.

Model Loading

- Loads a pre-trained LSTM model (text_lstm_emotion_model.pth) alongside a tokenizer and label encoder to handle text inputs for emotional detection.
- Leveraging a pre-trained CNN model (cnn_audio_model.h5) and a label encoder, audio model detects emotions based on audio characteristics.
- Using a pre-trained ResNet model (ResNet 18), video model analyzes emotions in video content by using rich video data retrieval.

- LLM (Ollama) interacts with the LLama3 model using the Ollama library, therefore allowing a thorough analysis of text, audio, and video data.

Preprocessing

- To prepare the input for the LSTM model, the text is preprocessed using a tokenizer and Keras pad_sequences.
- Using Librosa, extract Mel-frequency cepstral coefficients (MFCCs) from audio files; additionally, moviepy.editor allows one to extract audio from video files.
- Processes video frames using OpenCV (cv2) and gets them ready for feature extraction via the ResNet model with torchvision.transform.
- Using the speech_recognition library, audio material from videos may be transcribed into text so enabling text-based emotional analysis of spoken content.

Emotion Predicting

- Emotions are predicted from text using the loaded LSTM model.
- Derives emotional forecasts from audio features using the CNN model.
- Text, audio features, and video features extracted by the LLM are integrated into the video for identifying the most possible dominant emotion in a multimodal context.

User Interface (Streamlit)

- With a sidebar for simple navigation, layout offers a well-organized arrangement supporting text, audio, and video media inputs.
- Included in the "Introduction" section is a means of guiding users across the system.
- Clearly shows, depending on user inputs, expected feelings.
- File management controls file uploads and provides interim processing storage.
- This deployment highlights the possibility of efficiently detecting emotions by merging several machine learning models, especially LSTM and multimodal techniques. The development achieved with models such as LLama3 and their embeddings emphasizes the continuous requirement of ever more accurate and sophisticated emotion identification systems in the future by means of continual developments.

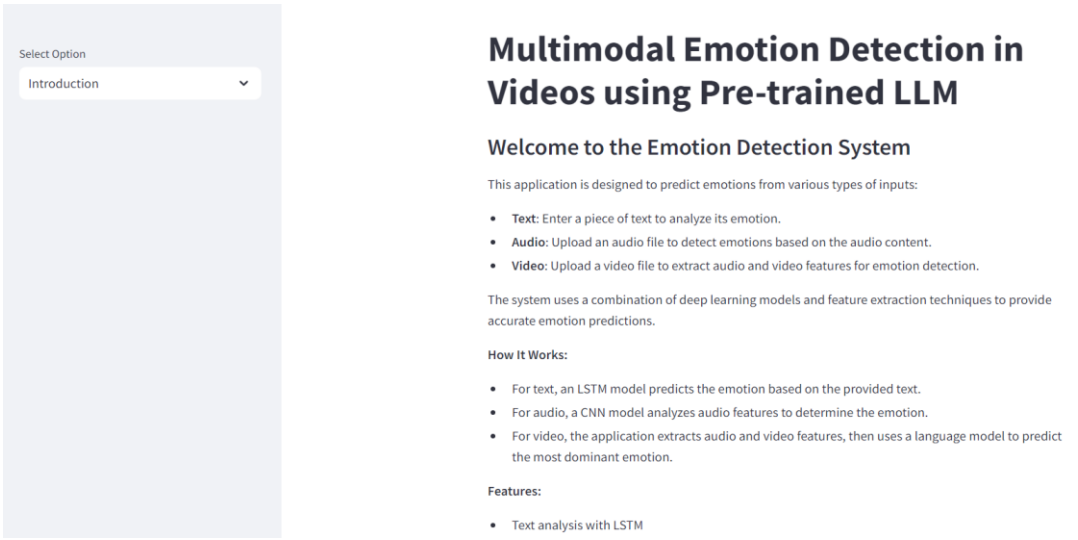

Fig 4.24 Introduction to application

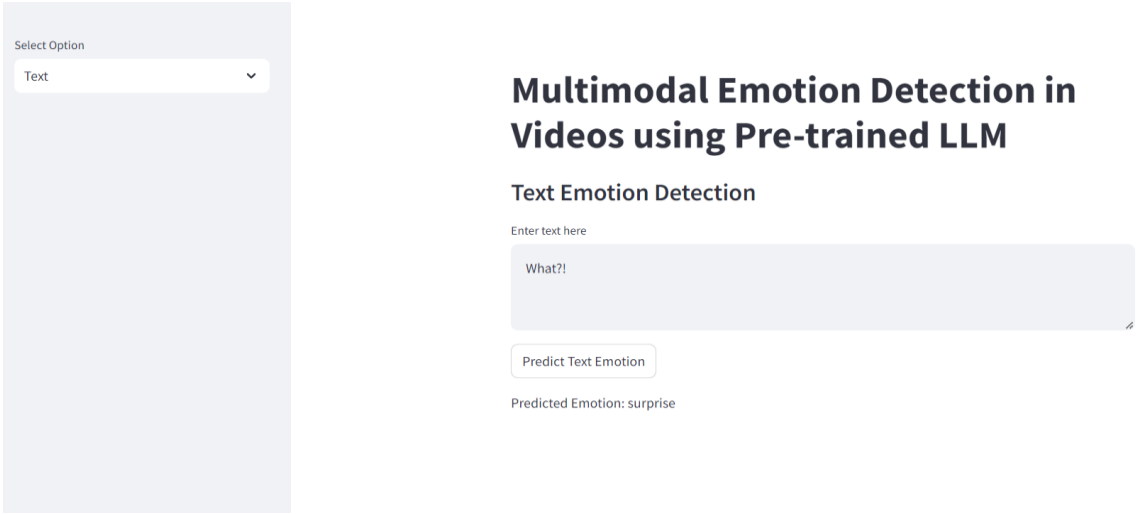

Fig 4.25 Emotion detection using Text data

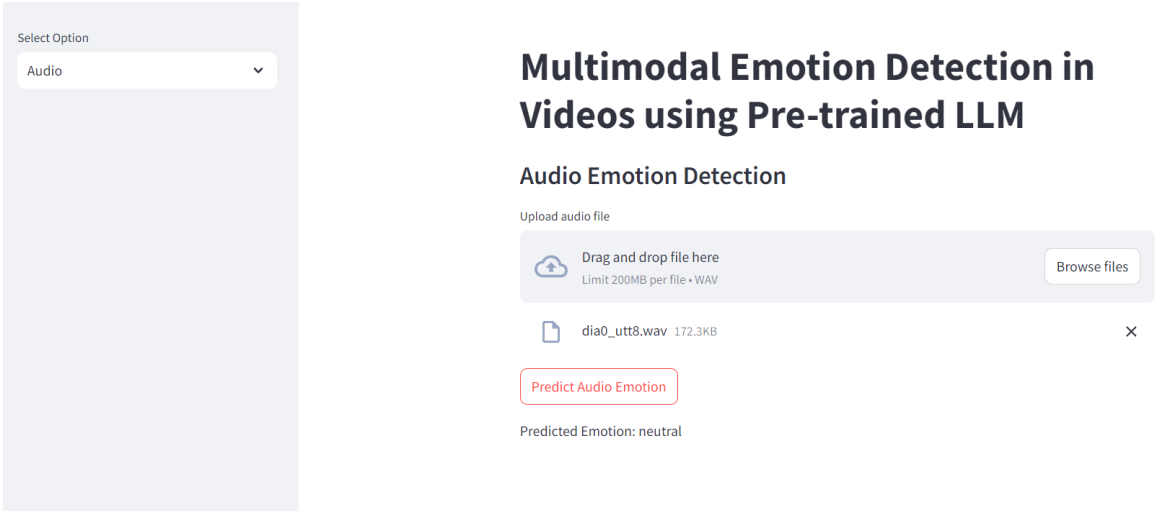

Fig 4.26 Emotion Prediction using Audio data

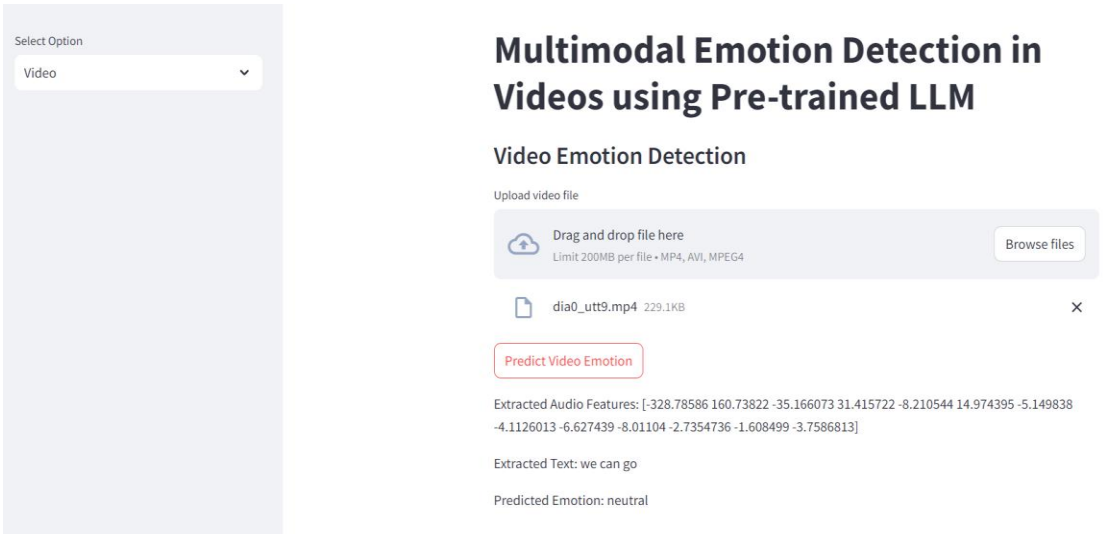

Fig 4.27 Emotion Prediction using Video data (Multimodal emotion prediction)

Explanation of Code

- Users can enter text, which the LSTM model then processes and forecasts for emotional state.
- Users can upload an audio recording, the system uses Librosa to extract MFCC features, then forecasts the emotion using a pre-trained CNN model.
- Users can upload a video file, and the system picks out the audio from it. Using voice recognition, the audio is handled to extract features and translate them into text. Video frames then are taken out and feature extracted using a pre-trained ResNet model. These elements text, audio, and video are aggregated into a prompt, and an LLM (large language model) forecasts the most dominating feeling from the film.

CHAPTER 5: CONCLUSION

5.1 Reintroducing Objectives

The study sought to evaluate the current multimodal techniques for emotion recognition by means of textual, visual, and audio data integration to improve the accuracy of emotional detection in videos. Aiming to fully assess the effectiveness of the multimodal emotion detection system, the research also concentrated on using a pre-trained Large Language Model (LLM)-based framework for multimodal data processing.

5.2 Implications of Results

The results show that the accuracy and efficacy of emotion identification in videos are much enhanced by including textual, visual, and auditory modalities. Improving the processing and analysis of multimodal data became especially successful with the use of a pre-trained LLM-based framework. This method not only enhanced the identification of intricate emotional signals but also showed possible uses in other disciplines like psychological research, video content analysis, and human-computer interaction.

5.3 Limitations

The study has constraints, including the difficulties with the computational complexity of processing and integrating multimodal data, which would impede scaling in practical uses. Dependency on pre-trained models also runs the danger of introducing prejudices that can compromise the generalizability of the outcomes. Furthermore, the availability of premium multimodal datasets presented a challenge that affected analytical resilience.

5.4 Lessons Learned

The project underlined the important need of including several data modalities to more precisely capture and analyze emotions in videos. It underlined the significance of choosing and adjusting pre-trained models to fit the requirements of multimodal emotion detecting applications. The study also underlined the requirement of sophisticated methods in this field

by offering understanding of the difficulties of coordinating and synchronizing several data streams.

5.5 Future Works

Investigating more effective models and approaches for multimodal data processing can help future studies to solve computational difficulties. Additionally, efforts should be directed toward reducing possible biases in pre-trained models and broadening the dataset to incorporate more diverse and practical situations. Furthermore, looking at the integration of other modalities, such physiological inputs, could improve the dependability and accuracy of emotional detecting systems even further.

CHAPTER 6: REFERENCES

Hans, A. S. A., & Rao, S. (2021, February). *A CNN-LSTM based deep neural networks for facial emotion detection in videos*. INTERNATIONAL JOURNAL OF ADVANCES IN SIGNAL AND IMAGE SCIENCES. https://doi.org/10.29284/ijasis.7.1.2021.11-20

Bhattacharya, P., Gupta, R. K., & Yang, Y. (2021, April). *Exploring the Contextual Factors Affecting Multimodal Emotion Recognition in Videos*. IEEE Xplore. https://doi.org/10.1109/TAFFC.2021.3071503

Heredia, J., Silva, E. L., Cardinale, Y., Amado, J. D., Dongo, I., Graterol, W., & Aguilera, A. (2022, February). *Adaptive Multimodal Emotion Detection Architecture for Social Robots*. IEEE Xplore. https://doi.org/10.1109/ACCESS.2022.3149214

Aslam, A., Sargano, A. B., & Habib, Z. (2023, June). *Attention-based multimodal sentiment analysis and emotion recognition using deep neural networks*. ScienceDirect. https://doi.org/10.1016/j.asoc.2023.110494

Middya, A. I., Nag, B., & Roy, S. (2022, March). *Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities*. ScienceDirect. https://doi.org/10.1016/j.knosys.2022.108580

Graterol, W., Diaz-Amado, J., Cardinale, Y., Dongo, I., Lopes-Silva, E., & Santos-Libarino, C. (2021, February). *Emotion Detection for Social Robots Based on NLP Transformers and an Emotion Ontology*. MDPI. https://doi.org/10.3390/s21041322

Elyoseph, Z., Refoua, E., Asraf, K., Lvovsky, M., Shimoni, Y., & Hadar-Shoval, D. (2023, November). *Capacity of Generative AI to Interpret Human Emotions From Visual and Textual Data: Pilot Evaluation Study*. JMIR Publications. https://doi.org/10.2196/54369

Han, B., Yoo, C.-H., Kim, H.-W., Yoo, J.-H., & Jang, J. (2023, June). *Deep Emotion Change Detection via Facial Expression Analysis*. ScienceDirect. https://doi.org/10.1016/j.neucom.2023.126439

An, Y., Lee, J., Bak, E., & Pan, S. (2023, August). *Deep Facial Emotion Recognition Using Local Features Based on Facial Landmarks for Security System*. Tech Science Press. https://doi.org/10.32604/cmc.2023.039460

Alkaabi, N., Zaki, N., Ismail, H., & Khan, M. (2022, November). *Detecting Emotions Behind the Screen*. MDPI. https://doi.org/10.3390/ai3040056

Manalu, H. V., & Rifai, A. P. (2024, February). *Detection of Human Emotions Through Facial Expressions Using Hybrid Convolutional Neural Network-Recurrent Neural Network Algorithm*. ScienceDirect. https://doi.org/10.1016/j.iswa.2024.200339

Bisogni, C., Cimmino, L., Marsico, M. D., Hao, F., & Narducci, F. (2023, June). *Emotion recognition at a distance: The robustness of machine learning based on hand-crafted facial features vs deep learning models*. ScienceDirect. https://doi.org/10.1016/j.imavis.2023.104724

Kalyta, O., Barmak, O., Radiuk, P., & Krak, I. (2023, August 31). *Facial Emotion Recognition for Photo and Video Surveillance Based on Machine Learning and Visual Analytics*. Applied Sciences. https://doi.org/10.3390/app13179890

Akhand, M. A. H., Roy, S., Siddique, N., Kamal, M. A. S., & Shimamura, T. (2021, April 27). *Facial Emotion Recognition Using Transfer Learning in the Deep CNN*. Electronics. https://doi.org/10.3390/electronics10091036

Mocanu, Tapu, & Zaharia. (2023, April). *Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning*. ScienceDirect. Retrieved May 2024, from https://doi.org/10.1016/j.imavis.2023.104676

Chen, S. Y., Hsu, C. C., Kuo, C. C., Ting-Hao, Huang, & Ku, L. W. (2018, February 23). *EmotionLines: An Emotion Corpus of Multi-Party Conversations*. arXiv.org. https://doi.org/10.48550/arXiv.1802.08379

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2018, October 5). *MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations*. arXiv.org. https://doi.org/10.48550/arXiv.1810.02508