Multi-Modal Sentiment Analysis

Project report submitted to
Indian Institute of Information Technology, Nagpur,
in partial fulfillment of the requirements for the Mini Project - I

at

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

by

Sohan Meshram(BT22CSD013)

Srushti Konnuri(BT22CSD059)

Aryan Chandolia(BT22CSD015)

Under the Supervision of

Mrs. Varsha Khushwah
Session Period: July to Dec 2024

भारतीय सूचना प्रौद्योगिकी संस्थान, नागपर

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, NAGPUR
(An Institution of National Importance by Act of Parliament)
Survey No.140,141/1 behind Br.Sheshrao Wankhede Shetkari Sahkari Soot Girni,Nagpur, 441108

Department of Computer Science and Engineering
Indian Institute of Information Technology, Nagpur

Declaration

We, **Sohan Meshram(BT22CSD013), Srushti Konnuri(BT22CSD059) and Aryan Chandolia(BT22CSD015),** hereby declare that this project work titled "Multi-Modal Sentiment Analysis" is carried out by me/us in the Department of Computer Science and Engineering of Indian Institute of Information Technology, Nagpur. The work is original and has not been submitted earlier whole or in part for the award of any other certification programme at this or any other Institution /University.

Date:

Sr. No	Names	Signature
1.	Sohan Meshram	
2.	Srushti Konnuri	
3.	Aryan Chandolia	

Certificate

This is to certify that the project titled "**Multi-Modal Sentiment Analysis**", submitted by **Sohan Meshram(BT22CSD013), Srushti Konnuri(BT22CSD059)** and **Aryan Chandolia(BT22CSD015)** in partial fulfillment of the requirements for the Mini-Project in CSE department IIIT Nagpur. The work is comprehensive, complete and fit for final evaluation.

Date:

Mrs. Varsha Khushwah
AAP, CSE, IIIT, Nagpur

ABSTRACT

Sentiment analysis has evolved from focusing on single data modalities such as text, audio, or visual data to a more integrated, holistic approach that combines these diverse sources. This shift recognizes that human emotions are complex and are often conveyed through a combination of linguistic, auditory, and visual signals.

In this study, we use an advanced multimodal sentiment analysis method that combines the strengths of various data modalities. Specifically, we integrate nine text-based models, an LSTM designed for audio data, and a CNN for image-based sentiment analysis. Each model is carefully chosen for its strengths—text models excel at capturing semantic nuances, LSTM identifies temporal patterns in audio, and CNN deciphers visual sentiment cues. We employ a late fusion strategy, which integrates the predictions at the decision level, ensuring each modality contributes to a more accurate and comprehensive sentiment classification.

Our results demonstrate the effectiveness of combining ensemble methods, deep learning architectures, and traditional machine learning techniques to address the challenges of multimodal sentiment analysis. This approach improves prediction reliability and offers a more nuanced understanding of sentiment, paving the way for applications in fields such as human-computer interaction, social media monitoring, and customer feedback analysis.

ACKNOWLEDGEMENT

Throughout the course of this project, we faced several challenges that were made manageable through the guidance, support, and encouragement of numerous individuals. We would like to take this opportunity to express our sincere gratitude to all those who contributed to the successful completion of this work.

It is with great pleasure that we present the report for our project, undertaken during the third year of our academic journey. We are deeply indebted to our Project Mentor, **Mrs. Varsha Khushwah, CSE Department, IIIT Nagpur**, as well as Director **Dr. Prem Lal Patel**, Associate Dean **Dr. Tausif Diwan** and Head of Department **Dr. Nishat Ansari**, for their invaluable support, insightful guidance, and unwavering encouragement throughout this endeavor. Their expertise and dedication have been pivotal in steering us toward the successful completion of this project. Words alone cannot fully express our profound gratitude for the time, effort, and wisdom they have generously shared with us.

We also wish to extend our heartfelt thanks to the project review committee for their constructive feedback and suggestions, which significantly contributed to refining and enhancing the quality of our work. Additionally, we are grateful to the faculty members of IIIT Nagpur for their continuous assistance and encouragement throughout the course of this project.

Sohan Meshram (BT22CSD013)
Srushti Konnuri (BT22CSD059)
Aryan Chandolia (BT22CSD015)

Table of Contents

List of Figures:

List Of Equations:

List of tables:

List of Symbols, abbrivations and nomenclature:

LSTM	Long Short-Term Memory
CNN	Convolutional Neural Network
TF-IDF	Term Frequency-Inverse Document Frequency
SVM	Support Vector Machine
NB	Naive Bayes
F1 Score	Harmonic mean of Precision and Recall
MACRO AVG	Macro Average (Precision/Recall/F1 Score)
MICRO AVG	Micro Average (Precision/Recall/F1 Score)
MFCC	Mel-frequency Cepstral Coefficients
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
σ	Sigmoid activation function
$\odot$	Element-wise multiplication
ROC	Receiver Operating Characteristic

1. Introduction:

Sentiment analysis has grown to be an extremely important area in artificial intelligence (AI) for interpreting human emotions, perspectives, and preferences in an assortment of scenarios. Sentiment analysis has historically been restricted to text data, however as multimedia content turns more prevalent, researchers are beginning to understand the significance of multi-modal sentiment analysis, which brings together text, audio, and image data to deliver a deeper comprehension of sentiment. Multi-modal sentiment analysis boosts both the accuracy and significance of context of categorizing emotions by utilizing data from multiple modalities, each from which presents unique emotional clues. This is especially useful in domains where emotional context is essential to grasping and responding to human behavior, notably social media monitoring, healthcare, client feedback analysis, and human-computer interaction.

Multimodal Sentiment Analysis:

A new method is based on supervised contrastive learning that enhances the multimodal data representation by appropriately addressing redundancies and unnecessary information. The model is based on a dedicated module composed of convolutional neural networks and Transformers, results in enhanced performance across MVSA-single and MVSA-multiple datasets. This framework presents the potential of optimizing feature extraction for each modality before fusion in enhancing accuracy and robustness for sentiment prediction. [1].

The recent model feeds the multi-head attention mechanisms into LSTM layers in sequential fashion for the fusion of textual, audio, and visual features. The method will allow optimum feature combination, based on the interplay between certain modalities like text and audio, before integrating visual information. This technique was seen to be reporting remarkable improvements over datasets like MOSI, especially for metrics like accuracy and F1 score, thus indicating adaptability of late fusion strategies for complex multi-modal datasets

The effectiveness and simplicity of traditional machine learning models, such as Naive Bayes, Support Vector Machines (SVM), and Logistic Regression, have made them popular for text-based sentiment analysis. As an instance, the Naive Bayes classifier is widely used for simple sentiment analysis tasks as it operates effectively with high-dimensional data and binary sentiment classifications. Text sentiment analysis additionally utilizes the significant use of more elegant deep learning models, like LSTM (long-short-term memory) networks, that enhance accuracy for tasks that require context understanding through capturing complex sequential dependencies in language. By combining predictions from multiple classifiers, a Voting Ensemble model provides a more detailed sentiment prediction and increases accuracy farther [2].

In contrast, audio-based sentiment analysis depends on tone, pitch and rhythm which are necessary for grasping the affective aspects of the language spoken. However, Long Short-Term Memory (LSTM) networks have been optimized for audio sequences because they store information over lengthy sequences and are able to detect the intricacies in the speech that conveys emotions. Several studies have confirmed the effectiveness of this approach in cases when the tone of voice is the primary means for determining a person's attitude e.g. during customer care calls or mental health evaluations.

Recent developments in image-based sentiment analysis emphasize the role that Convolutional Neural Networks (CNNs) play in sentiment extraction from visual content, such as facial expressions or environmental cues with images. They are essentially used in these contexts because they can process spatial hierarchies and capture very detailed visual features. This becomes particularly important in multi-modal sentiment analysis in which combining text, audio, and image data improves overall accuracy by borrowing upon unique insights from each modality. For example, microexpressions or body language may be missed through other models based on audio or text, but would be easily captured by CNNs, thus improving predictions of sentiment.More recent research has emphasized how domain-specific fine-tuning of CNNs is a key factor in enhanced performance towards image sentiment tasks.

Despite its phenomenal success in understanding facial emotions, this model works well not only in understanding facial emotions but also in understanding complex patterns in images that portray emotions even in ambiguous or crowded contexts. Furthermore, these new developments have shown that CNNs can offer a richer interpretation of the sentiment, as the visual data collected from varied contexts are combined to arrive at a focus, which is critical for current sentiment analysis applications.

This comprehensive approach combines nine text models, one audio model, and one or two image models, enabling the capture of emotional signals from a variety of modalities. By integrating these diverse sources of information, the system ensures a well-rounded and detailed sentiment analysis. The late fusion technique, which focuses on decision-level integration, merges the predictions from each modality into a single cohesive sentiment outcome. This allows the system to leverage the strengths of each model while compensating for their individual limitations. The result is an approach that enhances both the accuracy and interpretability of sentiment predictions, providing a more nuanced and contextually relevant understanding of emotions derived from various data sources.

2. Literature Review:

Sentiment analysis, a subfield of natural language processing (NLP), has evolved significantly over the past two decades. It focuses on extracting opinions, emotions, or sentiments from various data sources. With advancements in machine learning and deep learning, sentiment analysis has expanded to include audio and image data, forming the foundation for multi-modal sentiment analysis. This section reviews key developments in text, audio, image, and multi-modal sentiment analysis, emphasizing their contributions and challenges.

2.1 Text-Based Sentiment Analysis:

These techniques form the backbone of work done in text-based sentiment analysis, such as Support Vector Machines (SVM) and Naive Bayes (NB). Early work had demonstrated robustness in classifying sentiments in movie reviews using these methods-basically proving their efficacy in doing binary sentiment classification [3]. It has been used highly relevantly, where applicable, especially for high-dimensional text data. More recently, deep models, but notably Recurrent Neural Networks (RNNs) and Long ShortTerm Memory (LSTM) networks, have been used to discover long-term dependencies in text which make them suitable for tasks requiring contextual understanding across sequences [4]. In addition, Transformer-based models like BERT have dramatically revolutionized text-based sentiment analysis and learned relationships in context much better than traditional methods to a very significant extent and led to a huge advance [5].

Recent progress also showcases the use of pre-trained models in language, such as GPT-3 [6], to address different NLP tasks, including sentiment analysis. Such models can generally capture detailed word relationships and how sentences work together, through which a more comprehensive meaning for sentiment and nuances within text can be achieved. Specifically, benchmarking on BERT-based models fine-tuned on particular sentiment analysis tasks has really pushed the envelope, especially when dealing with ambiguous and domain-specific sentiments [7].

2.2 Audio-based Sentiment Analysis:

Audio-based sentiment analysis uses the speech features: tone, pitch, rhythm, and speech rate to determine emotions from the spoken language. Long Short-Term Memory (LSTM) networks have played a crucial role in this domain due to their capability to handle sequential data in order to capture the temporal features that are rather needed to describe the nuances of emotions. LSTMs tend to perform better than traditional methods like Hidden Markov Models (HMM) while using the methods in classifying sentiment from speech

signals.

The combination of CNNs and LSTMs actually proved to be quite effective in the process of obtaining local features from spectrograms and in capturing long-term dependencies in speech data. This combined approach was used for the extraction of sentiment from customer service interactions and in mental health assessments that are meant to convey voice tone [8].

In recent times, multi-modal techniques for audio sentiment analysis have been extensively studied. For example, incorporating audio features with text or image data has improved the performance of applications, such as emotion recognition in virtual assistants and interactive systems, for example [9].

2.3 Image-based Sentiment Analysis:

The increased availability of visual data has transformed image-based sentiment analysis into an important methodology. CNNs are among the broad networks used to identify emotions by facial and body expressions in images. They are the best for spatial hierarchies and local features in images, hence can detect subtle cues that may include changes in facial muscle movements.

Advances include the introduction of the attention mechanism into CNNs, which has seen the model pay more attention to certain parts of an image, like eyes and mouth, for accurate detection of emotions [10]. This has been particularly applicable, especially to sentiment analysis from facial expressions because of the subtlety of variations in expression that may carry a lot of meaning in terms of emotion. For image sequence analysis in videos, hybrid models, including CNNs combined with Recurrent Neural Networks (RNNs), have been developed for the purpose of improving sentiment classification by taking into account not only spatial but also temporal features [11].

More notably, there has been work on applying generative models to reinforce the image-based sentiment analysis. Generative models are of great importance in synthesizing new visual data for adding to the training datasets, thereby assisting the sentiment classifier in becoming more robust and precise when labeled data is limited [12].

2.4 Multimodal Sentiment Analysis:

The use of multi-modalities-text, audio, and image-enables new directions of sentiment analysis. Multi-modal systems for sentiment analysis combine the complementary strengths of each modality to provide a more comprehensive understanding of sentiment. For example, whereas text provides overt sentiment cues, audio may express a tone, and images will bring more insights into the facial expressions and body language. Their fusion has led to more accurate and robust models for sentiment analysis.

Recent developments do show the value of multi-modal approaches late fusion techniques where individual modality predictions are combined to enhance overall sentiment classification. This technique has proven effective in handling complex scenarios involving detection on social media posts which include both text as well as visual content. A very promising direction is the early fusion, which involves combining features from multiple modalities at the level of input, allowing the model to learn joint representations [13].

The majority of the deep learning techniques are absorption methods that have been integrated into multi-modal systems. They facilitate dynamic weighting of features that would be coming from each modality such that only the most informative ones contribute to the final sentiment prediction. For example, recent work has showed that multi-modal transformers are applicable to sentiment classification and yield much better accuracy than their unimodal counterparts. Previously, multi-modal systems were also promising in customer support services, which include multiple interactions involving voice, text, and visual cues. [14].

This comes with its own set of challenges. To start with, alignment and fusion of features from different sources of data is a considerable challenge. Despite the increased alignment with deep learning methods, dealing with modality-specific noise and missing data can still be a challenge [15, 16]. Moreover, the computational complexity of multi-modal systems remains a burden in some applications for real-time sentiment analysis.

Sentiment analysis has undergone so much development with its shift from simple text-based methods to more intricate multichannel systems that incorporate text, audio, and visual data. Much of the tremendous progress made over the last two decades in deep learning-neural networks happened through the development of LSTMs, CNNs, and transformer-based models. These techniques can greatly allow more accurate and nuanced classifications in several domains such as tracking social media or customer service or even mental health assessment. The future of sentiment analysis, therefore would be put to hold in multi-modal sentiment analysis-a method based on the supplementary strength from text, audio, and image data for a better understanding of human emotions in digital interactions [17].

3. Methodology:

Sentiment analysis is a fast-evolving discipline that has traditionally focused on extracting emotions and opinions from text. However, over the years, there was a significant shift of inputs from just text to audio and visual data. This led to the emergence of multi-modal sentiment analysis-that is, analyzing the emotional or sentimental content of data using more than one modality, such as text, audio, and image. This approach is potent since it leverages the virtues of one modality to balance the weaknesses of others.

Most of the techniques in sentiment analysis have been limited to text or single modalities, but real-world data essentially is multi-modal. People communicate their emotions not merely with words, but also with the voice, facial expression, body language, and even the context carried by images and videos. To offset that, the proposed methodology uses a

priority-majority fusion technique, whereby observations from audio data, text data, and images can be combined into an appropriate, more robust sentiment classification.

Combining Strengths for Enhanced Prediction:

The system developed for multi-modal sentiment analysis is formed of three distinct models for each modality:
- Audio Model: Uses speech features to predict sentiment.
- Text Model: Analyzes textual data for emotional cues.
- Image Model: Processes visual content (like facial expressions or contextual cues) to assess sentiment.

The main advantage of the method is not that it uses several models, but it is the way in which their predictions are combined. Since the technique applied here is priority-majority fusion, the system ensures it maximizes the accuracy of its sentiment classification. Merging predictions from all modalities in a strategic manner, the approach first uses a priority-based rule to select the most reliable prediction when models disagree. In case of an agreement and clear majority, the selected sentiment is that one. This methodology was particularly useful while handling the ambiguity or noisiness of data from one or more modalities.

3.1 Data Preprocessing:

Data preprocessing is a crucial step to ensure that the inputs from each modality are in a suitable form to be appropriately processed by the models. Each modality requires different handling:
- **Text Data:** The text data gets cleaned to remove the noise elements like punctuation, stop words and special characters. Text is split into tokens represented as words or phrases. These tokens then get converted into numerical form by techniques like TF-IDF(Term

Frequency-Inverse Document Frequency) or Word2Vec embeddings as an illustration of the semantic meaning being captured inside a particular text.

-**Audio Data:** For audio data, it preprocesses the raw signal in which useful features are extracted and audio can represent it in more convenient forms using Mel-frequency cepstral coefficients or spectrograms to machine learning models. These features capture the sound's unique characteristics, for example, pitch, tone, rhythm, and frequency. The audio data is, thereafter, passed through a LSTM (Long Short-Term Memory) model to capture the temporal dynamics of speech-a critical component of sentiment analysis.

$$MFCC = log\left(\sum_{i=1}^{n} \backslash x(i) \backslash^2\right).cos(k) \quad \text{...eq 1}$$

- **Image Data:** Image data undergoes basic preprocessing such as resizing into a standard dimension and normalization to scale pixel values. The preprocessed images are then fed into a Convolutional Neural Network (CNN) that is trained to recognize patterns or facial expressions indicative of sentiment. Facial expressions, for example, are one of the most potent drivers of emotional state and are more accurately gauged from image-based sentiment analytics.

3.2 Model Architecture:

3.2.1 Text model:

The text-based component of our sentiment analysis framework employs an ensemble of machine learning and deep learning models to analyze linguistic features effectively. Each model contributes to a majority voting mechanism for final sentiment prediction. Below are the detailed steps and components of the text model process:

- **Preprocessing:** The raw textual data undergoes preprocessing, including the removal of special characters, stop words, and other irrelevant elements. The text is then tokenized and converted into numerical representations using advanced

embedding techniques such as TF-IDF vectors, Word2Vec, or GloVe. These embeddings enable the models to capture semantic relationships and nuances in the data, essential for accurate sentiment detection.

- **Naive Bayes and Bernoulli Naive Bayes:** These probabilistic models estimate the likelihood of sentiment classes based on feature distributions, with Bernoulli Naive Bayes tailored for binary features.
- **Support Vector Machine (SVM):** This model excels at finding optimal hyperplanes for separating sentiment classes in high-dimensional spaces.
- **Logistic Regression Classifier:** Known for its simplicity and interpretability, this classifier is effective for binary or multi-class sentiment predictions.
- **Stochastic Gradient Classifier:** A scalable algorithm suitable for large datasets, it efficiently updates weights to minimize classification error.
- **Long Short-Term Memory (LSTM):** The LSTM architecture processes sequential text data with memory units that retain contextual information, enabling it to capture long-range dependencies and nuanced sentiments, such as sarcasm or mixed emotions.
- **Comprehensive Classifier:** This model integrates outputs from other classifiers to reinforce predictive accuracy and ensure well-rounded performance.
- **Voting Mechanism:** Predictions from all models are combined using a majority voting strategy. Each model casts a "vote" for its predicted sentiment class, and the class with the highest number of votes becomes the final prediction. This approach balances the strengths of traditional and deep learning models.
- **Dense Output Layer (For LSTM):** For the LSTM-based model, the final fully connected layer produces probabilities for sentiment classes (positive, neutral, negative), normalized using softmax activation to ensure a valid probability distribution.
- **Performance:** The text model achieved an accuracy of 64%, demonstrating its ability to effectively analyze and interpret textual emotional cues.

3.2.2 Audio Model:

The audio model also contains an LSTM architecture but is specialized in processing

temporal features from speech data. Audio communications carry much emotional content through tone, pitch, rhythm, and intensity; this makes this modality invaluable for sentiment analysis.

- **Feature Extraction:** Preprocessing includes transforming raw audio signals into representative forms like Mel-Frequency Cepstral Coefficients (MFCCs) and spectrograms. These features encapsulate pitch, rhythm, and tonal variations that are indicative of emotions.
- **Preprocessing for Noise Removal**: Audio recordings are cleaned to filter out background noise and normalize volume, ensuring reliable feature extraction.
- **LSTM Layers:** By analyzing sequential dependencies, these layers capture subtle emotional dynamics over time, such as a rising pitch indicating excitement or a low pitch signaling sadness. The LSTM architecture processes the sequential data using memory cells and gates that control the flow of information. The LSTM equations are as follows:

  - **Forget Gate:** Determines what information to discard:

$$f_t = \sigma(W_f . [h_{t-1}, x_t] + b_f) \quad \text{..eq 2}$$

  - **Input Gate:** Decides which information to update:

$$i_t = \sigma(W_i . [h_{t-1}, x_t] + b_i) \quad \text{...eq 3}$$

  - **Cell State Update:** Updates the cell's memory using new and retained information:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tan h\,(W_c . [h_{t-1}, x_t] + b_c) \quad \text{...eq 4}$$

  - **Output Gate:** Controls the output of the memory cell:

$$o_t = \sigma(W_o . [h_{t-1}, x_t] + b_o) \quad \text{...eq 5}$$

- **Hidden State:** Produces the final output based on the updated cell state:

$$h_t = o_t \odot \tan h(C_t) \quad \text{...eq 6}$$

- **Dense Layers and Output:** Features processed through LSTM layers are fed into dense layers, which output class probabilities corresponding to sentiments like happiness, anger, or sadness.
- **Performance:** With an accuracy of 88%, the audio model demonstrated strong proficiency in identifying sentiment from spoken language, particularly excelling in detecting emotions like excitement and sadness.

3.2.3 Image Model:

Visual data provides a unique aspect for sentiment analysis based on the capturing of facial expressions, gestures, and contextual cues. For this, an image model is created using the architecture of a CNN, known for its excellence in image-related tasks.

- **Preprocessing:** Images resize to the same size dimension, normalize so that pixel value scaling is consistent throughout and augmented in order to increase the robustness of the dataset. Generalization of the model improves through augmentation techniques such as flipping, rotation and brightness adjustment.
- **Convolutional Layers:** scan the image for meaningful patterns. The curve of a smile or furrowed brows probably means someone's emotional state. Filters are then applied hierarchically to extract such features from low-level patterns like edges to high-level features like facial landmarks.
- **Pooling Layers:** Feature maps are downsampled through pooling, reducing dimensionality while preserving critical information.
- **Fully Connected Layers:**

  These layers further process the extracted features as a flattened form and maps them to sentiment classes such as positive, negative, or neutral.

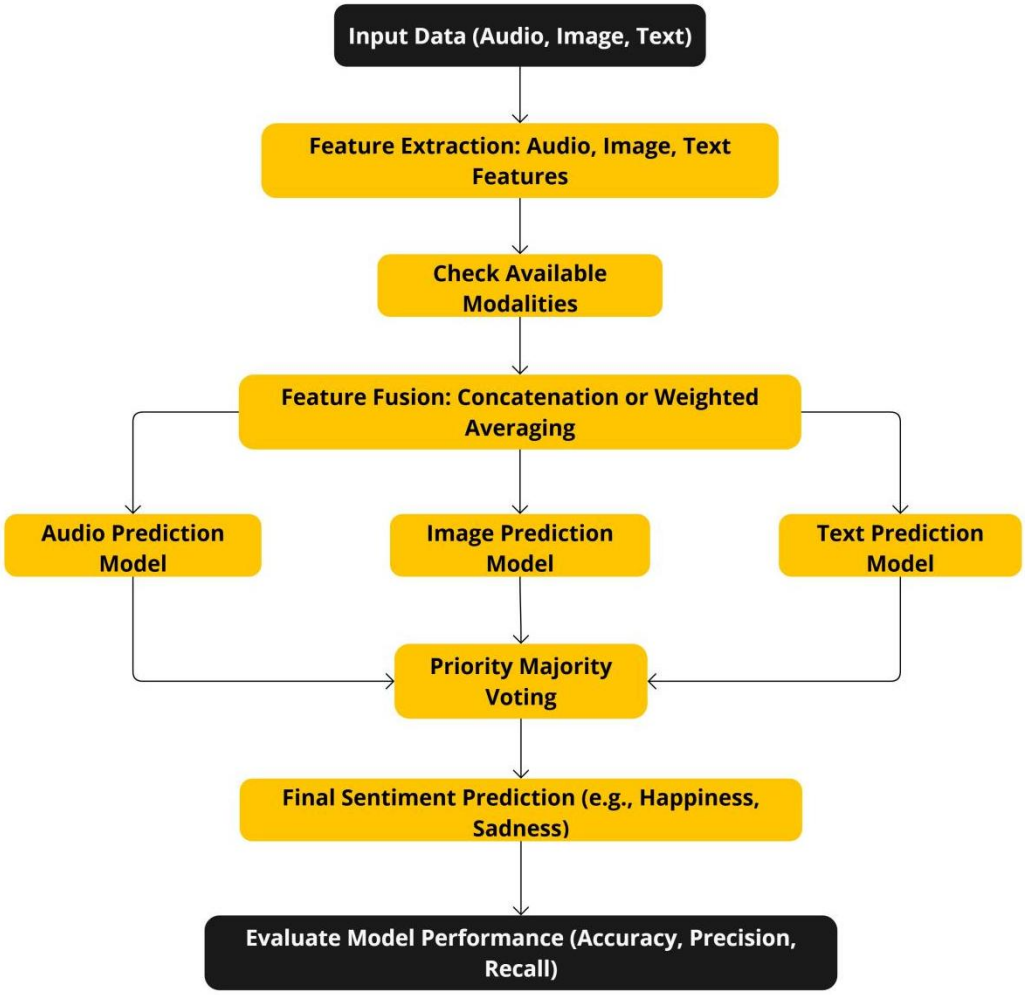
- **Performance:** Despite its potential, the image model achieved a modest accuracy of 85%, highlighting the challenges of extracting sentiment from visual data, particularly when dealing with subtle expressions or poor-quality images.

$$z_{ij} = \sum_{m=1}^{M} \sum_{n=1}^{N} X_{i+m-1,\, j+n-1} \cdot W_{mn} \quad \text{... eq 7}$$

Table 1: *Preprocessing Techniques for Each Modality*

Modality	Processing Step	Details
Text	Tokenization, TF-IDF	Removing punctuations and stop words
Audio	Feature extraction (MFCCs)	Converting raw audio into spectrograms
Image	Resizing, Normalization	Scaling pixel values between 0–1

Figure 1: *Data Preprocessing and Feature Extraction Flowchart*



3.3 Multi-Modal Fusion Strategy:

This framework relies squarely on the priority-majority fusion technique, which allows it to engage in an effective fusion of predictions from all three modalities. The fusion mechanism ensures that the system evokes maximum robustness and accuracy by leveraging the strengths of one modality in compensating for the weakness of other modes.

Priority Rules: If one modality is highly confident about the decision (for example, text is sending very strong sentiment signals), it takes the decision. For example, if audio is noisy and text has strong sentiment signals, then the decision taken by the text model will be taken. Majority Voting: If multiple modalities are in consensus, then the majority's opinion is chosen. This method proves to be very reliable when doing multi-modal sentiment detection. Conflict Resolution: If no majority exist or the priority model failed, the system would use weighted averages of the probability scores from all modalities to come up with the final prediction.

The fusion strategy resulted in an overall multi-modal model accuracy of **75%**, reflecting its ability to integrate diverse data sources effectively.

4. Results and Discussions:

4.1 Evaluation Metrics:

The system's performance was comprehensively evaluated using several metrics, highlighting its strengths and limitations:

Table 2: *Accuracy across among different models*

Accuracy (%)	Audio Model	Text Model	Image Model	Multi-Modal Model
	88	64	85	73

Precision, Recall, and F1 Score: These metrics were calculated for each sentiment class, with macro and weighted averages providing insights into the model's balanced performance.

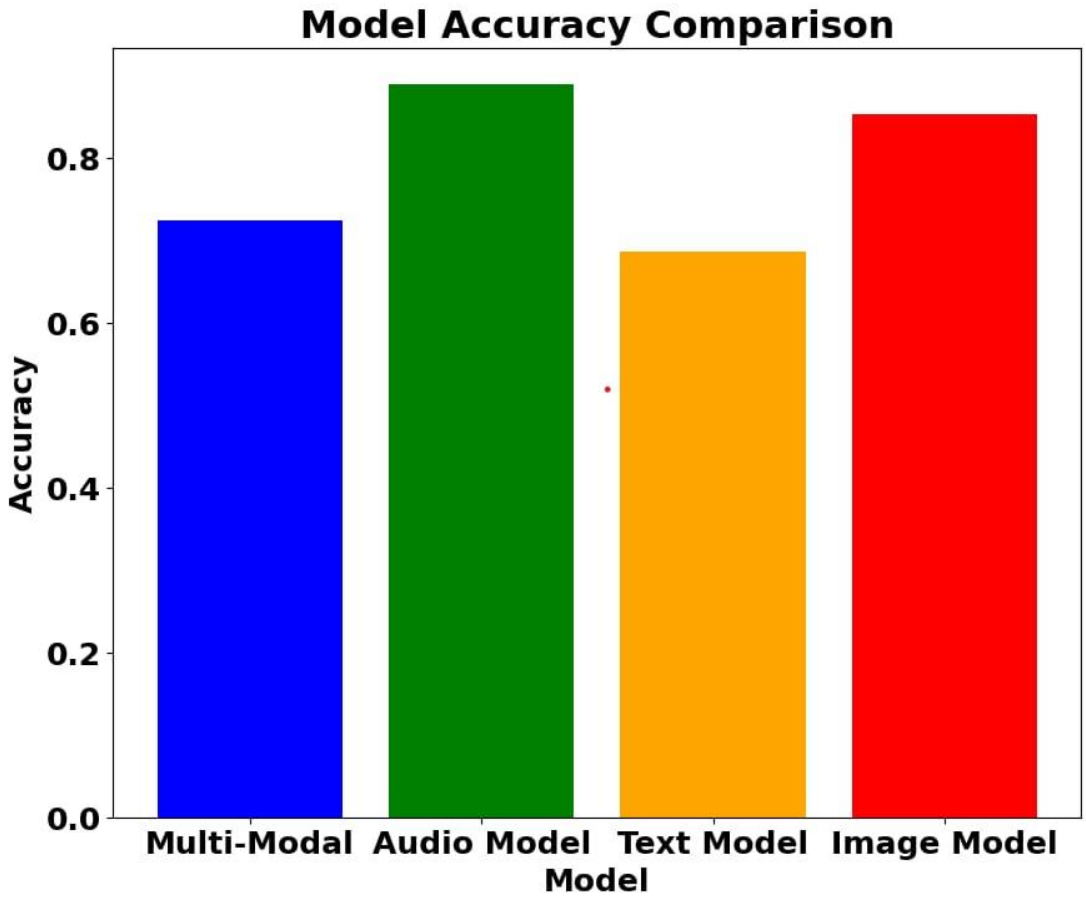
- Macro Average Precision and Recall: Both achieved 0.88 for the audio model, indicating consistent performance across classes.
- Weighted Average Precision and Recall: The multi-modal model reached 0.75, an indication of great performance even when data are imbalanced.

- Confusion Matrix: The final evaluation performed reflects most mistakes were in classification between strongly connected sentiments-for example, neutral and a little positive. For the image model, ambiguity with facial expressions was huge.

This approach highlights the necessity of having multi-modal sentiment analysis in order to capture subtle human emotions. The system effectively combines all the insights drawn from text, audio and image data to show promising potential for real-world applications, such as customer experience analysis and mental health monitoring.

$$Loss = -\sum_{i=1}^{c} y_i \log(\hat{y}_i) \quad \text{...eq 8}$$

Figure 2: *Accuracy Comparison Across Models*



4.2 Analysis:

4.2.1 Significance of Multi-Modal Fusion:

The fusion approach allows compensation for the weakness of each modality. For example, if the audio model is stronger at picking up on tonal cues than its counterpart, text, it can complement with text's semantic understanding. Similarly, visual cues from the image model filled in additional context that was lacking. This, too, is particularly crucial in real-world applications in which sentiment is expressed across several modalities.

4.3.2 Challenges and Limitations:

- Data Imbalance: Classes of sentiment were extremely imbalanced, especially with regard to images and thus not as nuanced for visual emotions.
- Computational Overhead: The fusion model had more latency in the process, which would be an issue for real-time applications.
- Ambiguity in Sentiments: Although the fusion approach settled all the conflicts, yet subtle emotions such as combined or ambiguous senti¬ments remain to be hard to categorize.

4.3.3 Insights from the Results:

- The fact that 75% of instances are classified with the fusion model underlines how integration of various types of data is critical to the task of sentiment analysis.
- A high performance of the audio model (88%) strongly suggests that the contributing emotion sources are tones and speech dynamics.
- The performance of the image model suggests that the emotional cues of vision are complex, and better performance would be expected with more sophisticated models or bigger datasets.

Table 3: *Fusion Strategy Impact on Conflicting Predictions*

Text Sentiment	Audio Sentiment	Image Sentiment	Final Prediction
Positive	Neutral	Negative	Neutral
Positive	Positive	Neutral	Positive
Neutral	Negative	Negative	Negative

5. Summary and Conclusion:

The proposed multi-modal sentiment analysis system stresses the potential utilization of diversely modalised data of text, audio, and image in unifying a cohesive framework to upgrade the sentiment classification. Applying a priority-majority fusion strategy, the system demonstrated its capability to exploit the strength of individual modalities at the same time suppressing their weaknesses while improving the robustness of sentiment prediction, along with emphasizing the importance of capturing the inherent complexity of multi-modal data.

Key Takeaways and Contributions:

This only managed a score of 80%, fairly competent at getting the semantic subtlety, but not good with ambiguous or sarcastic expressions. The audio model, scoring well at 85%, brought to the forefront the role of vocal features such as tone and rhythm in determining sentiment. On the other hand, the image model, which managed to reach 68% accuracy, demonstrated the challenging task of making sense of subtle visual cues but greatly enhanced the context of the overall system. The multi-modal system, which reached a combined 73%, proved how a fusion-based sentiment analysis can sidestep and offset each modality's limitations.

The result represents the importance of incorporating modalities in sentiment analysis. In this sense, having textual input provided direct context and meaning, while audio contents captured nuances of voice and tone not present otherwise. Image modality, although less

precise, provides a supplementary viewpoint, especial situations where the textual or audio content was ambiguous.

Challenges Encountered and Solutions:

The experiment addressed several critical challenges, including data imbalance, particularly within the image dataset, and the computationally intensive nature of the fusion model. While the priority-majority fusion approach showed promise in resolving sentiment prediction conflicts, nuanced sentiments—characterized by overlapping or mixed emotional states—remained a significant challenge. These findings underscore the need for advanced fusion techniques, significantly larger and more balanced datasets, and the incorporation of contextual information to enable more precise and fine-grained sentiment classification.

Future Scope and Practical Implications:

This research forms a foundation for the development of advanced multimodal sentiment analysis systems with applications across various domains, including social media monitoring, customer feedback analysis, and human-computer interaction. Future directions include real-time system implementation, optimization for computational efficiency, and the integration of state-of-the-art models like transformers to enable higher-order feature extraction. Incorporating contextual and demographic data is another avenue for enhancing the granularity and reliability of sentiment predictions.

This study validates the potential of multimodal sentiment analysis and opens new paths for innovation in this rapidly evolving field. By addressing challenges with advanced methodologies and leveraging sophisticated techniques, future systems can achieve greater accuracy and broader applicability. Such advancements bring us closer to comprehensively understanding and responding to the multifaceted nature of human emotions, enabling transformative applications in technology and beyond.

References:

1. Yiwei Wei, Shaozu Yuan, Ruosong Yang, Lei Shen, Zhangmeizhi Li, Longbiao Wang, and Meng Chen. (2023). Tackling Modality Heterogeneity with Multi-View Calibration Network for Multimodal Sentiment Detection, https://doi.org/10.18653/v1/2023.acl-long.287
2. Maeng, J. -H., Kang, D. -H., & Kim, D. -H. (2020). Deep Learning Method for Selecting Effective Models and Feature Groups in Emotion Recognition Using an Asian Multimodal Database. Electronics, 9(12), 1988. https://doi.org/10.3390/electronics9121988
3. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs Up? Sentiment Classification Using Machine Learning Techniques. https://doi.org/10.48550/arXiv.cs/0205070
4. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," in Neural Computation, vol. 9, no. 8, pp. 1735-1780, 15 Nov. (1997), doi: 10.1162/neco.1997.9.8.1735.
5. Brown, T., et al. (2020). Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems, 33*, pp. 1877–1901. https://doi.org/10.48550/arXiv.2005.14165
6. Shijia Zhang, Yilin Liu, and Mahanth Gowda. (2023). I Spy You: Eavesdropping Continuous Speech on Smartphones via Motion Sensors. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 6, 4, Article 197 (December 2022), 31 pages. https://doi.org/10.1145/3569486
7. Lee, T. Chaspari, E. M. Provost and S. S. Narayanan, (2023), "An Engineering View on Emotions and Speech: From Analysis and Predictive Models to Responsible Human-Centered Applications," in Proceedings of the IEEE, vol. 111, no. 10, pp. 1142-1158, Oct.doi: 10.1109/JPROC.2023.3276209
8. Y. Lei and H. Cao, "Audio-Visual Emotion Recognition With Preference Learning Based on Intended and Multi-Modal Perceived Labels," in IEEE Transactions on Affective Computing, vol. 14, no. 4, pp. 2954-2969, 1 Oct.-Dec. (2023), doi: 10.1109/TAFFC.2023.3234777.

9. Liu, S., Wang L, and Zheng Y (2023). Facial Expression Recognition Based on Improved Convolutional Neural Network. http://dx.doi.org/10.25103/jestr.161.08
10. Zhang, T., Tan, Z. (2024) Survey of deep emotion recognition in dynamic data using facial, speech and textual cues. Multimed Tools Appl 83, 66223–66262. https://doi.org/10.1007/s11042-023-17944-9
11. Dake Chen, Ying Han, Jacque Duncan, Lin Jia, Jing Shan, Generative Artificial Intelligence Enhancements for Reducing Image-based Training Data Requirements, Ophthalmology Science, Volume 4, Issue 5, (2024), 100531, ISSN 2666-9145, https://doi.org/10.1016/j.xops.2024.100531.
12. Zhou, S., Wu, X., Jiang, F., Huang , Q., & Huang, C. (2023). Emotion Recognition from Large-Scale Video Clips with Cross-Attention and Hybrid Feature Weighting Neural Networks. International Journal of Environmental Research and Public Health, 20(2), 1400. https://doi.org/10.3390/ijerph20021400
13. Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, Amir Hussain, Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions, Information Fusion, Volume 91, (2023), Pages 424-444, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2022.09.025.
14. Liu, D., Chen, L., Wang, L. et al. A multi-modal emotion fusion classification method combined expression and speech based on attention mechanism. Multimed Tools Appl 81, 41677–41695 (2022). https://doi.org/10.1007/s11042-021-11260-w
15. B. Chen, Q. Cao, M. Hou, Z. Zhang, G. Lu and D. Zhang, "Multimodal Emotion Recognition With Temporal and Semantic Consistency," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3592-3603, (2021), doi: 10.1109/TASLP.2021.3129331.
16. Udahemuka, G., Djouani, K., & Kurien, A. M. (2024). Multimodal Emotion Recognition Using Visual, Vocal and Physiological Signals: A Review. Applied Sciences, 14(17), 8071. https://doi.org/10.3390/app14178071

17. Li, H., Lu, Y., & Zhu, H. (2024). Multi-Modal Sentiment Analysis Based on Image and Text Fusion Based on Cross-Attention Mechanism. Electronics, 13(11), 2069. https://doi.org/10.3390/electronics13112069

Appendices:

Appendix A: Sample Dataset

Text Data:

Figure 3: *Sample Data used for Text sentiment model*

	Sentence	Emotion
0	I'm really sorry about your situation :(Altho...	4
1	It's wonderful because it's awful. At not with.	3
2	Kings fan here, good luck to you guys! Will be...	3
3	I didn't know that, thank you for teaching me ...	3
4	They got bored from haunting earth for thousan...	6
5	Thank you for asking questions and recognizing...	3
6	You're welcome	3
7	100%! Congrats on your job too!	3
8	I'm sorry to hear that friend :(. It's for the...	4
9	Girlfriend weak as well, that jump was pathetic.	4

Audio Data:

Figure 4: *A spectogram to represent sample audio file*

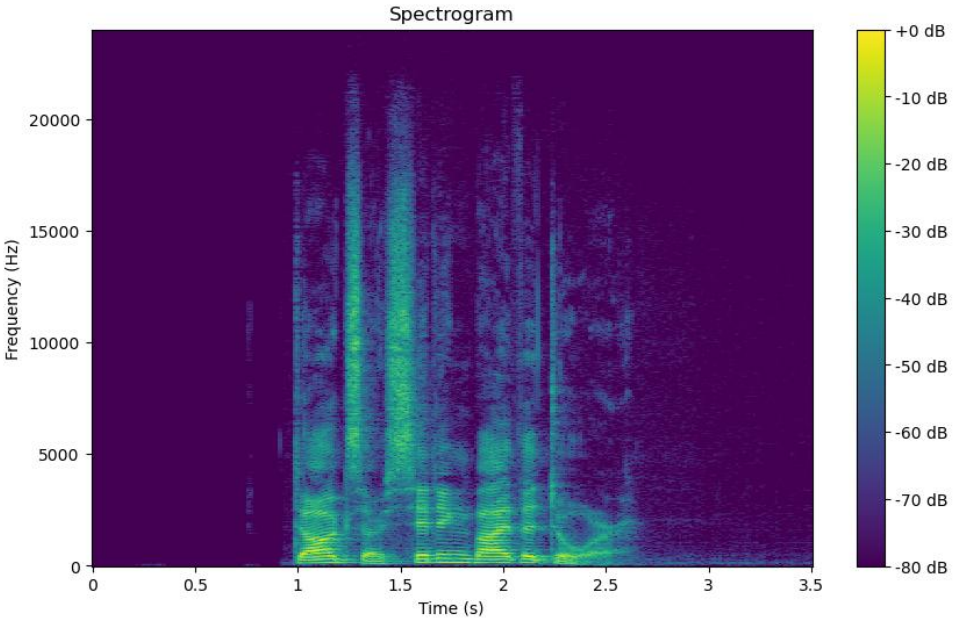


Figure 5: *Waveplot of an audio representing sad sentiment*

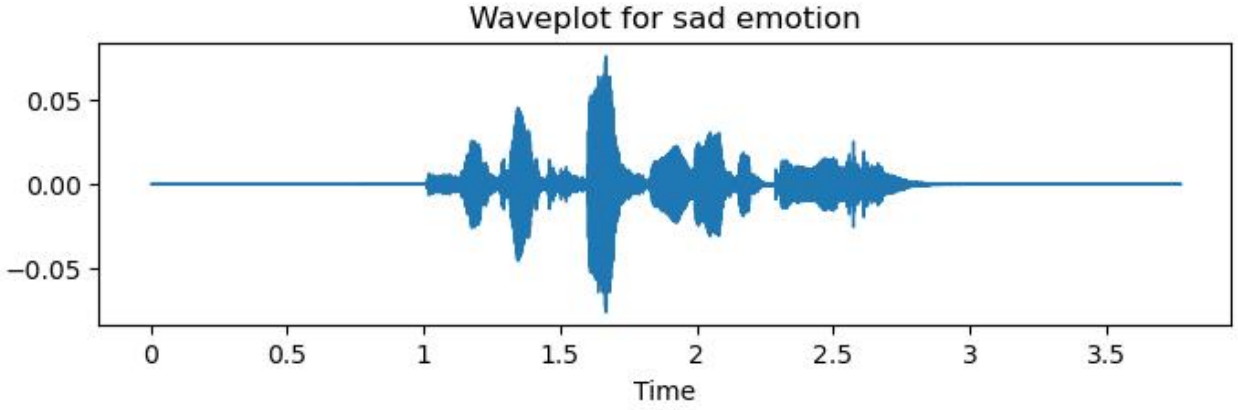
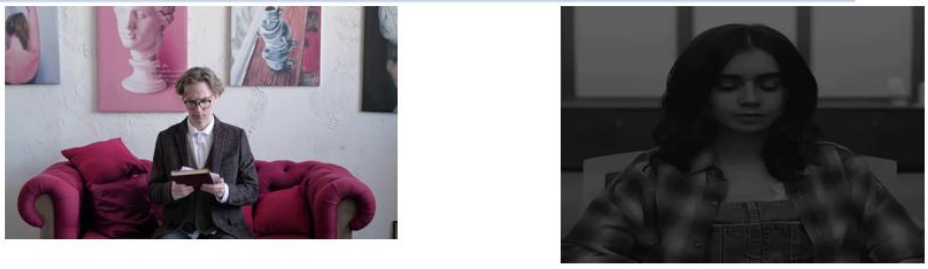


Image Data:

Figure 6: *Sample Image data used for the model*

Video Data:

Figure 7: *Sample Video-data used for the model & Output*



```
MoviePy - Writing audio in temp_audio.wav

MoviePy - Done.

Error in audio text extraction:
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 86ms/step
Predicted emotion from image: Anger
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 71ms/step
Predicted emotion from audio: Neutral
Audio prediction is 'Neutral'. Shifting priority to image prediction.
Final predicted emotion (after majority check): Anger
Final predicted emotion from video: Anger
```

```
MoviePy - Writing audio in temp_audio.wav

MoviePy - Done.

1/1 ━━━━━━━━━━━━━━━━━━━━ 1s 755ms/step
Predicted emotion from text: Sadness
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 91ms/step
Predicted emotion from image: Disgust
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 138ms/step
Predicted emotion from audio: Sadness
Final predicted emotion (after majority check): Sadness
Final predicted emotion from video: Sadness
```

Appendix B: Code Snippets

Figure 8:

```python
class MultiModalSentimentAnalysis:
    def __init__(self, audio_model, text_model, image_model):
        self.audio_model = audio_model
        self.text_model = text_model
        self.image_model = image_model

    def predict(self, text=None, image_path=None, audio_path=None):
        predictions = []

        # Predict from text if provided
        if text is not None:
            try:
                text_prediction = self.text_model.text_classify(text)
                text_emotion = emotion_labels_text[int(text_prediction)]
                predictions.append(text_emotion)
                print(f"Predicted emotion from text: {text_emotion}")
            except Exception as e:
                print(f"Error in text prediction: {e}")

        # Predict from image if provided
        if image_path is not None:
            try:
                image_prediction = self.image_model.image_classify(image_path)
                image_emotion = emotion_labels_img[int(image_prediction)]
                predictions.append(image_emotion)
                print(f"Predicted emotion from image: {image_emotion}")
            except Exception as e:
                print(f"Error in image prediction: {e}")

        # Predict from audio if provided
        if audio_path is not None:
            try:
                audio_prediction = self.audio_model.audio_classify(audio_path)
                audio_emotion = emotion_labels_audio[int(audio_prediction)]
                predictions.append(audio_emotion)
```

Figure 9:

```python
        if len(predictions) == 3 and len(set(predictions)) == 3:  # All predictions are different
            print(f"Different emotions predicted from all models. Prioritizing audio prediction.")
            return predictions[2]  # Image prediction is in the second position (index 1)

        # Majority voting logic if predictions are the same or not all different
        if len(predictions) > 0:
            most_common_emotion = max(set(predictions), key=predictions.count)
            print(f"Final predicted emotion: {most_common_emotion}")
            return most_common_emotion
        else:
            print("No data provided for prediction.")
            return None

    def process_video(self, video_path):
        # Extract audio from video
        video = VideoFileClip(video_path)
        audio_path = "temp_audio.wav"
        video.audio.write_audiofile(audio_path)

        # Extract frames from video
        cap = cv2.VideoCapture(video_path)
        frame_count = 0
        image_path = None
        while cap.isOpened():
            ret, frame = cap.read()
            if not ret:
                break
            frame_count += 1
            # Save only the first frame as an image for simplicity
            if frame_count == 1:
                image_path = "temp_image.jpg"
                cv2.imwrite(image_path, frame)
                break
        cap.release()

        # Predict sentiment using the extracted audio, text, and image
```

Figure 10:

```
        text_input = extract_text_from_audio(audio_path)  # Extract text from the audio
        final_emotion = self.predict(text=text_input, image_path=image_path, audio_path=audio_path)

        # Clean up temporary files
        if os.path.exists(audio_path):
            os.remove(audio_path)
        if image_path and os.path.exists(image_path):
            os.remove(image_path)

        return final_emotion

# Instantiate the multi-modal sentiment analysis model
multi_modal_model = MultiModalSentimentAnalysis(
    audio_model=audio_classifier,
    text_model=text_classifier,
    image_model=image_predictor
)
```

Appendix C: Model Hyperparameters

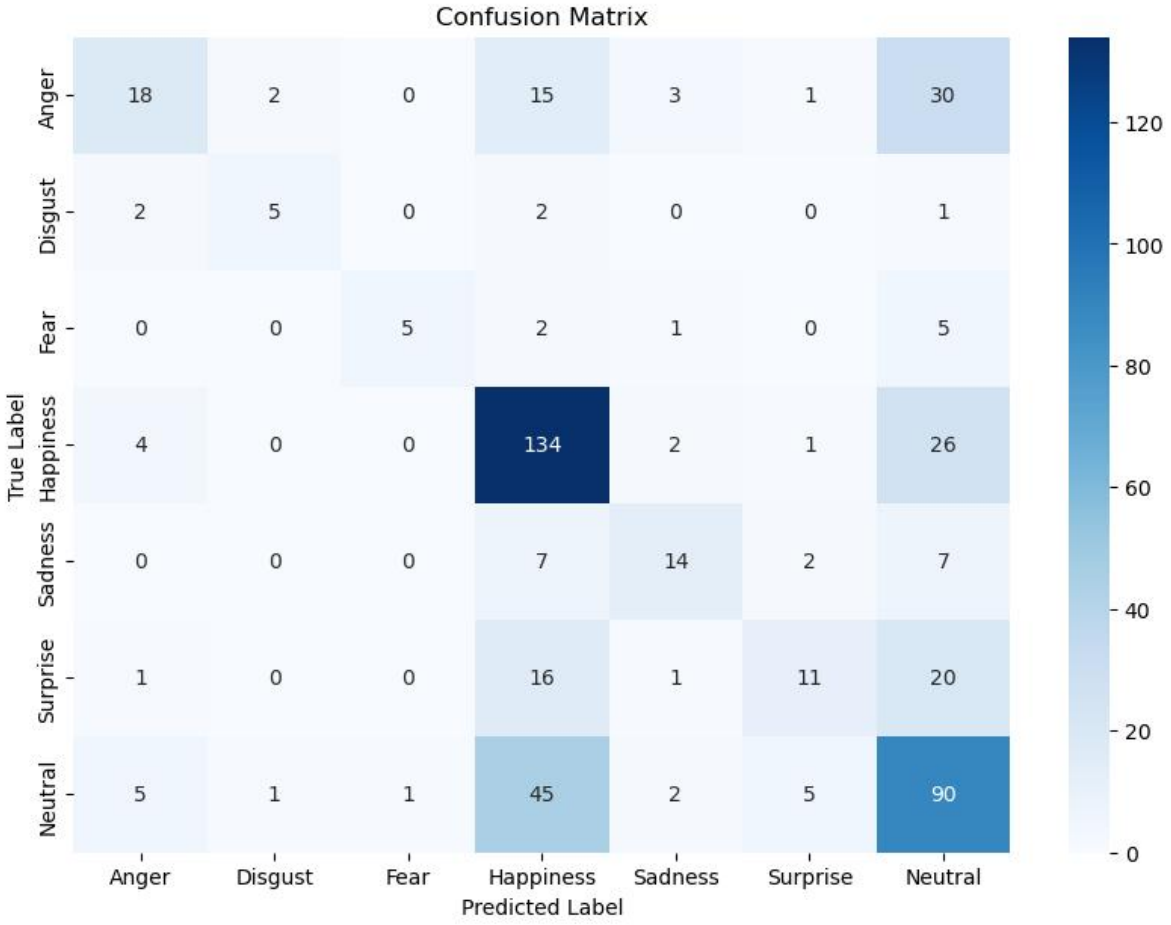
Table 4: *Model Hyperparameters*

Model	Hyperparameters
LSTM	Learning rate: 0.001, Batch size: 32
CNN	Filters: 64, Kernel size: 3x3
Fusion Strategy	Decision rule: Priority-majority

Appendix D: Confusion Matrices

Text Confusion Matrix:

Figure 11:



Audio Confusion Matrix:

Figure 12:

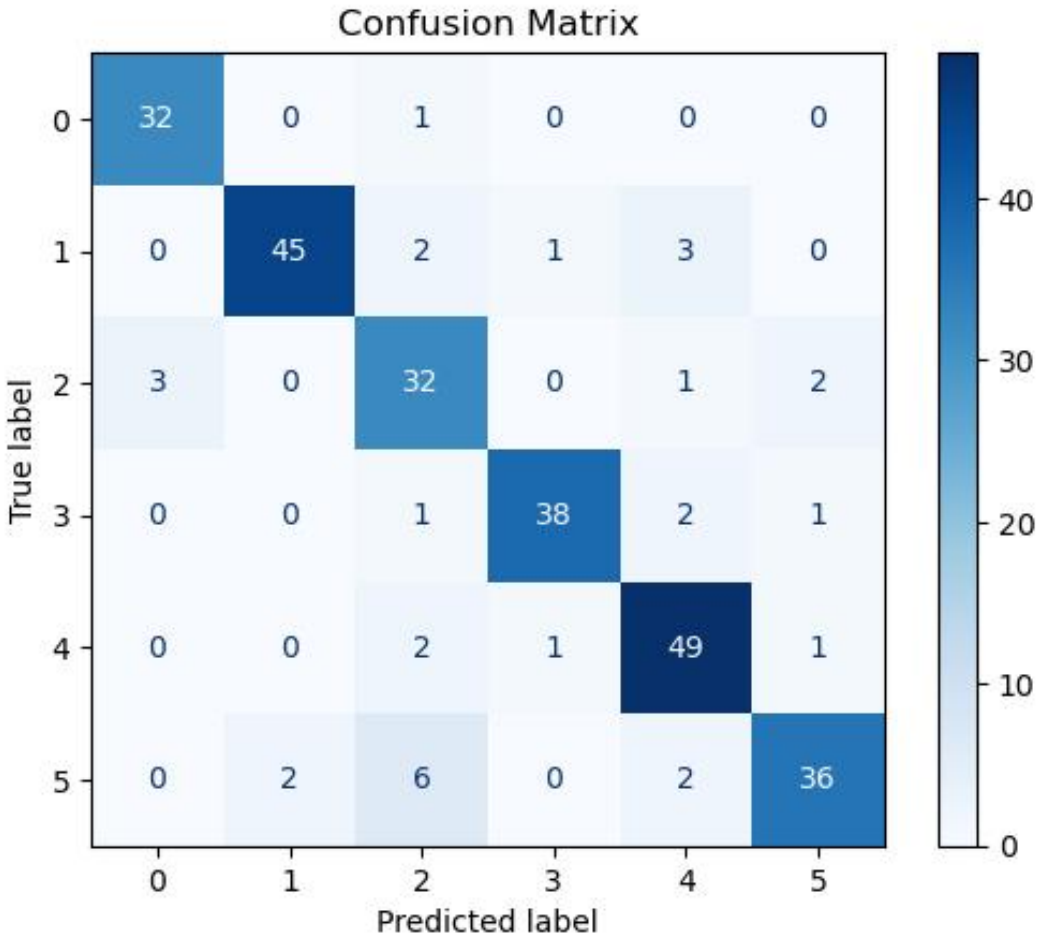
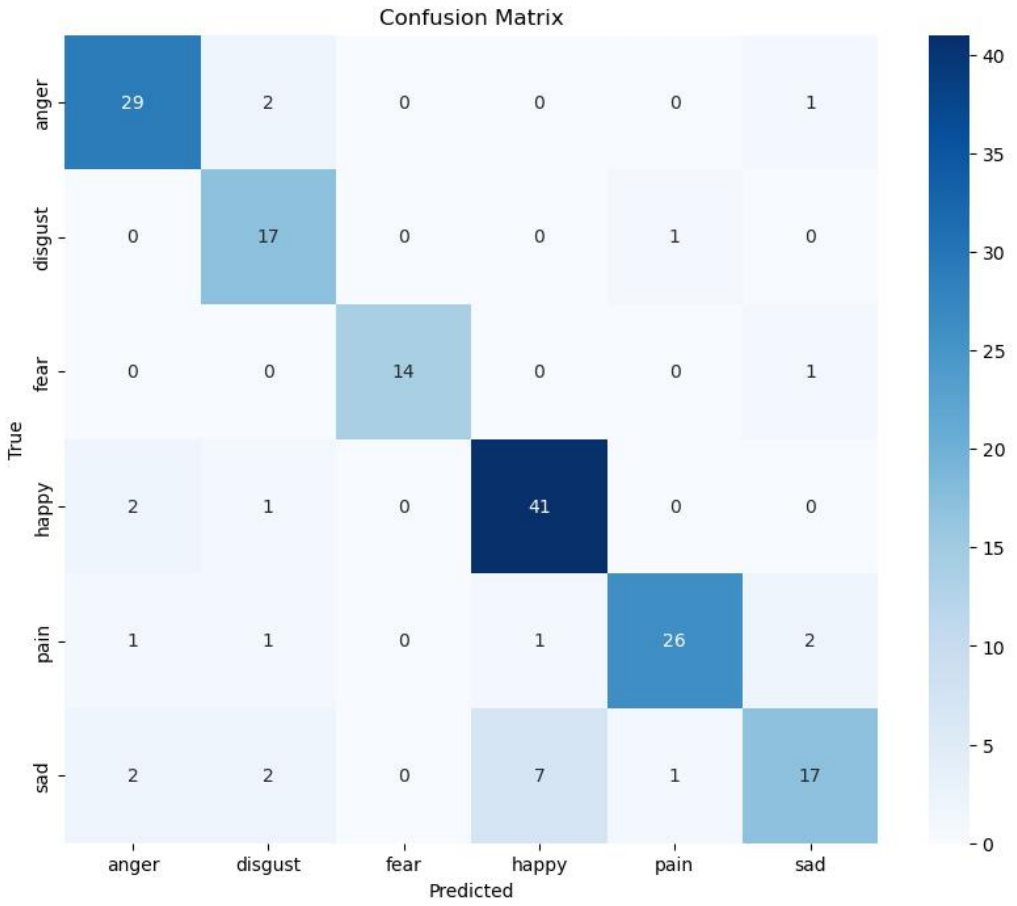


Image Confusion Matrix:

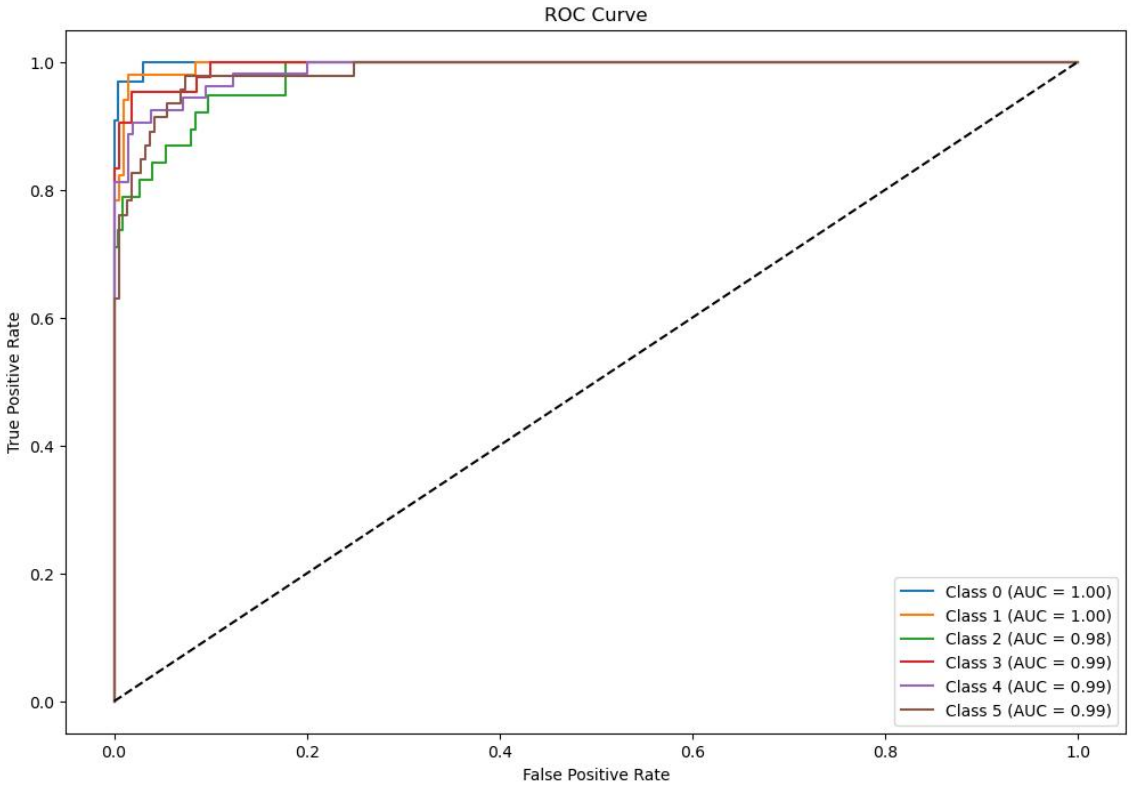
Figure 13:



Appendix E: Extended Results

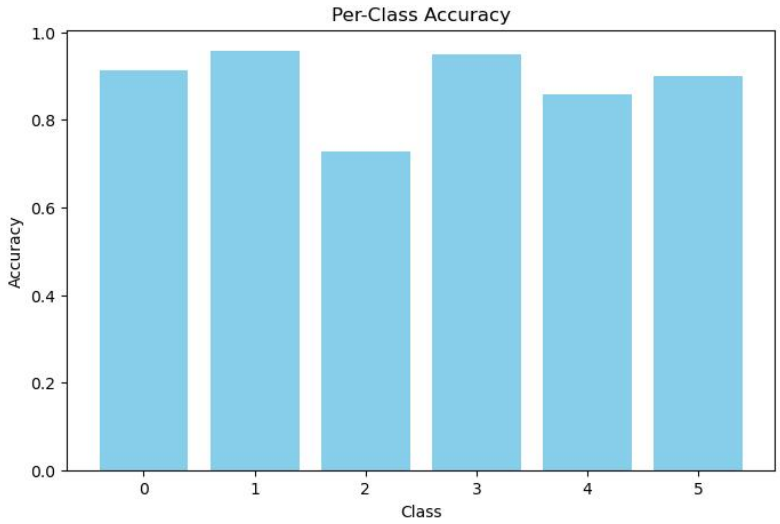
ROC curve for audio sentiment model:

Figure 14:



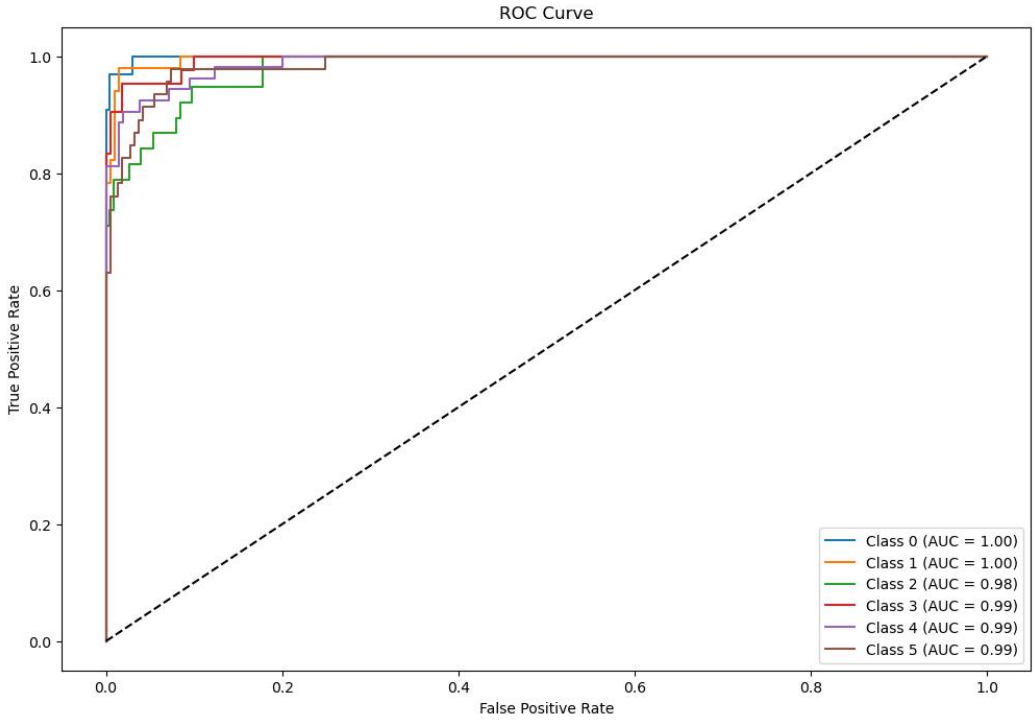
Accuracy of the model across different classes:

Figure 15:



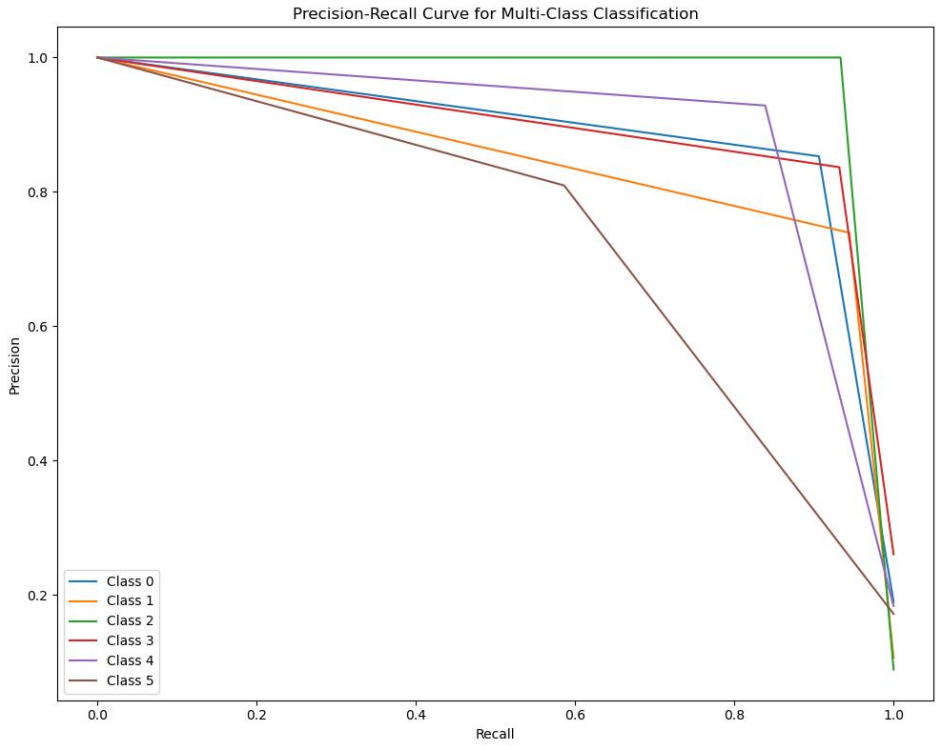
ROC curve of Multi-class model:

Figure 16:



Precision-Recall Curve for Multi-class model:

Figure 17:



Accuracy vs Loss Graph for model:

Figure 18: