Department of Electrical and Computer Engineering
North South University

Directed Research

Multimodal Sarcasm Detection in Bengali Language

Md Nazmus Saquib Khan **ID 2011537042**

Md Zian Raian **ID 2011394042**

Abid Hasan **ID 1921700042**

Kazi Nafisur Rahman **ID 2013628642**

Faculty Advisor:

Dr. Mohammad Abdul Qayum

Assistant Professor

ECE Department

Fall, 2025

LETTER OF TRANSMITTAL

Fall, 2025

To

Dr. Mohammad Abdul Motin

Chairman,

Department of Electrical and Computer Engineering

North South University, Dhaka

Subject: Submission of Directed Research Report on “Multimodal Sarcasm Detection in Bengali”

Dear Sir,

With due respect, we would like to submit our Directed Research Report on “Multimodal Sarcasm Detection in Bengali Language” as a part of our BSc program. The report deals with the development of a system that detects sarcasm in Bengali using multimodal data sources. This project was very valuable to us as it allowed us to explore the challenges involved in sarcasm detection and apply deep learning techniques to a language-specific problem.

We will be highly obliged if you kindly receive this report and provide your valuable feedback.

Sincerely Yours,

Md Nazmus Saquib Khan
ECE Department
North South University, Bangladesh

Md Zian Raian
ECE Department
North South University, Bangladesh

Abid Hasan
ECE Department
North South University, Bangladesh

Kazi Nafisur Rahman
ECE Department
North South University, Bangladesh

APPROVAL

Md Nazmus Saquib Khan (ID # 2011537042), Md Zian Raian (ID # 2011394042), Abid Hasan (ID # 1921700042) and Kazi Nafisur Rahman (ID # 2013628642) from Electrical and Computer Engineering Department of North South University, have worked on the Directed Research Project titled "Multimodal Sarcasm Detection in Bengali Language" under the supervision of Dr. Mohammad Abdul Qayum for partial fulfillment of the requirement for the degree of Bachelors of Science in Engineering and has been accepted as satisfactory.

Supervisor's Signature

…….

Dr. Mohammad Abdul Qayum

Assistant Professor

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

Chairman's Signature

…….

Dr. Mohammad Abdul Matin

Professor & Chair

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

DECLARATION

This is to declare that this project is our original work. No part of this work has been submitted elsewhere partially or fully for the award of any other degree or diploma. All project related information will remain confidential and shall not be disclosed without the formal consent of the project supervisor. Relevant previous works presented in this report have been properly acknowledged and cited. The plagiarism policy, as stated by the supervisor, has been maintained.

Students' names & Signatures

1. Md Nazmus Saquib Khan

_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

2. Md Zian Raian

_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

3. Abid Hasan

_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

4. Kazi Nafisur Rahman

_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

ACKNOWLEDGEMENTS

The The authors would like to express their heartfelt gratitude to their project supervisor, Dr. Mohammad Abdul Qayum [MAQm], Assistant Professor, Department of Electrical and Computer Engineering, North South University, Bangladesh, for his invaluable support, precise guidance, and advice pertaining to the experiments, research, and theoretical studies carried out during the course of this project.

Furthermore, the authors would like to thank the Department of Electrical and Computer Engineering, North South University, Bangladesh for facilitating the research. We would also like to thank our friends and families for their continual support.

ABSTRACT

Multimodal Sarcasm Detection in Bengali Language

Sarcasm detection in natural language is a challenging task, particularly when dealing with languages like Bengali, which have unique linguistic and cultural features. This project focuses on the development of a multimodal sarcasm detection system for Bengali using text, audio, and video modalities. By leveraging state-of-the-art models such as BERT for text, Wav2Vec and AST for audio, and ViT for video, the study aims to improve sarcasm detection by combining the strengths of these different data sources. The dataset used in this research consists of 739 video clips, labeled as sarcastic or non-sarcastic, with corresponding text transcriptions, audio recordings, and video files. Various configurations of MLP and attention-based models were explored to integrate the multimodal features and detect sarcasm. Results show that three-modality models significantly outperform two-modality setups, with MLP-based models yielding the best performance in terms of accuracy and F1 score. The findings highlight the importance of incorporating audio and video features alongside text for effective sarcasm detection and demonstrate the potential for multimodal models in emotion and sentiment analysis. The research also contributes to the Bengali language by providing a custom sarcasm detection dataset and establishing a foundation for further advancements in multimodal NLP.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

1.1 Background and Motivation

Sarcasm detection is a challenging and important task in the field of Natural Language Processing (NLP), speech processing, and computer vision. Sarcasm, which is a form of verbal irony, is often used to express criticism, humor in a manner that can be difficult to interpret. In traditional communication, sarcasm is conveyed through tone of voice, facial expressions, and gestures, alongside the verbal content. However, sarcasm detection in written or spoken language is difficult due to the subtlety of the cues that indicate sarcasm, making it one of the most complex aspects of sentiment analysis and emotion detection.

While sarcasm is easily understood by humans in everyday conversations, machines struggle to capture it. This challenge is compounded when analyzing textual data alone, as sarcastic statements often contradict their literal meaning. For instance, a sentence like "Oh, great! Another meeting!" can be interpreted as either a positive remark or a sarcastic one, depending on the context and tone.

As modern communication increasingly involves digital platforms, such as social media, chatbots, and voice assistants, the need for automated sarcasm detection systems is growing exponentially. Detecting sarcasm can improve the quality of interactions in these platforms by enabling machines to understand user's emotions, preferences, and sentiments more accurately. A deeper understanding of sarcasm can significantly enhance user experience in customer service systems, social media monitoring, and emotional AI applications.

In recent years, there has been a growing interest in multimodal machine learning, which involves integrating multiple sources of data, such as text, audio, and video, to improve machine learning performance. Multimodal sarcasm detection has proven to be more effective than relying on a single modality, as it combines different forms of information that provide complementary insights into the sarcasm in communication. The fusion of textual sentiment, speech tone, and visual cues can provide a richer understanding of the context, leading to more accurate sarcasm detection.

Despite the advancements in multimodal sarcasm detection, much of the existing research has been focused on Western languages, with limited resources and models available for Bengali, a language spoken by millions of people, primarily in Bangladesh and India. The scarcity of labeled sarcasm data for Bengali has led to a significant research gap, making it a valuable area for exploration. Additionally, the challenge of sarcasm detection is added more through the linguistic and cultural nuances specific to Bengali, which are distinct from those in English or other widely researched languages.

This project aims to bridge this gap by developing a multimodal sarcasm detection system in Bengali, using text, audio, and video modalities to improve the accuracy of sarcasm detection. By using state-of-the-art models for text, audio, and video feature extraction, such as BERT for text, Wav2Vec and AST for audio, and ViT for video, the project seeks to enhance the performance of sarcasm detection in Bengali by combining these multiple modalities in a unified framework. This study contributes to the growing field of multimodal emotion recognition by exploring the unique challenges of sarcasm detection in the Bengali language.

1.2 Purpose and Goal of the Project

The primary purpose of this project is to develop a multimodal sarcasm detection system for the Bengali language using text, audio, and video data. The goal is to create a system that can accurately detect sarcastic statements in Bengali, a task that has been largely underexplored in the field of multimodal emotion recognition for non-English languages.

The detection of sarcasm in communication—whether written, spoken, or visual—requires an understanding of both the literal meaning and the contextual cues that reveal the speaker's true intent. Sarcasm often relies on contradictory cues from different modalities: the text may appear positive, while the tone of the speaker’s voice and their facial expressions may indicate sarcasm..

In the context of Bengali, the complexity of sarcasm detection is heightened by linguistic subtleties and cultural context. The use of sarcasm in Bengali speech can differ from its use in Western languages, not just in phrasing, but in intonation patterns and non-verbal cues. As such, the project's aim is to investigate how Bengali speakers use sarcasm, and how these features can be effectively captured using modern deep learning techniques for natural language processing, speech recognition, and computer vision.

The specific goals of the project are as follows:

1. Development of a Multimodal Sarcasm Detection System: By integrating three modalities text, audio, and video the system aims to improve sarcasm detection accuracy by leveraging the complementary information from each modality. This multimodal approach enables the system to capture a wide range of sarcastic cues, whether they come from linguistic features, speech patterns, or visual expressions.

2. Exploration of Model Architectures:Tranformer based models were used in this project. The project explores both MLP-based models and attention-based models for combining features from different modalities. The goal is to understand how these architectures affect the model's performance and to determine whether the attention mechanism offers an advantage over traditional MLP-based fusion techniques.

3. Training and Evaluation on Bengali Sarcasm Data: Given the lack of publicly available Bengali sarcasm datasets, one of the key contributions of this project is the compilation of a Bengali sarcasm dataset from video clips, audio files, and text annotations. This dataset will serve as a benchmark for training and evaluating the model's performance.

4. Comparison of Modalities: By analyzing the contribution of each modality (text, audio, and video) in sarcasm detection, the project aims to identify the most important modality or combination of modalities for detecting sarcasm in Bengali. Additionally, it will investigate whether certain combinations of features, such as text and audio or text and video, offer more accurate results than using two modality.

Ultimately, the goal of this project is to improve the state of sarcasm detection in the Bengali language and to contribute to the broader field of multimodal sentiment analysis. By successfully developing a model that can effectively identify sarcasm across multiple modalities, this project will help lay the foundation for emotion-aware AI systems capable of engaging in more natural, human-like interactions with Bengali-speaking users. Additionally, the results could serve as a benchmark for future research in multilingual sarcasm detection, facilitating further development in languages with limited resources for sarcasm detection.

1.3 Organization of the Report

Chapter 2 reviews prior work: text-only sarcasm detection, prosodic and visual cues, multimodal datasets, fusion methods, temporal modeling vs. summarization, and deployment best practices. It also explains where our approach fits in this literature. Chapter 3 explains the methodology: data layout and aggregation, modality intersection, model design, training setup, metrics, complexity, and choices for reproducibility. Chapter 4 presents experiments: dataset details, training behavior, validation peaks, final test results, and error and threshold analyses, ending with practical steps to close the validation–test gap. Chapter 5 concludes with key findings, limits, and future work.

Chapter 2 Research Literature Review

2.1 Existing Research and Limitations

2.1.1 Text-only Sarcasm Detection in Bengali

Automated sarcasm detection has long posed a significant challenge in natural language processing. Because sarcasm often relies not just on the meaning of words but on contextual, prosodic, and visual cues, many researchers have argued that multimodal approaches — integrating text, audio, and visual information — can outperform purely textual methods, as demonstrated in the paper "A Survey of Multimodal Sarcasm Detection" by Farabi et al. [2]. One of the pioneering efforts in this direction was MUStARD, introduced in the paper "Towards Multimodal Sarcasm Detection (An Obviously Perfect Paper)" by Castro et al. (2019), which compiled a corpus of audiovisual utterances from television shows, annotated for sarcasm, along with contextual dialogue history [4]. Their initial experiments demonstrated that combining modalities could reduce the relative error rate substantially — showing that nonverbal cues significantly aid sarcasm detection.

Following this, more sophisticated multimodal fusion techniques have been proposed. For instance, a recent architecture described in the "Multi-modal sarcasm detection based on Multi-Channel Enhanced Fusion model" by Fang et al. (2024) seeks to maximize cross-modal information flow by using a bipolar semantic attention mechanism to capture inconsistencies across modalities, which are often characteristic of sarcasm [3]. Similarly, recent survey work summarizing the multimodal sarcasm detection literature, such as "Enhancing Multimodal Sarcasm Detection with Context-Aware Self-Attention Fusion and Word Weight Calculation" by Xue et al., argues that audio‑visual‑textual integration is increasingly important, especially for detecting sarcasm in realistic settings such as social media and dialogues [8].

A more recent approach, AMuSeD (2024), extends this by using data augmentation (on text and audio) combined with attentive fusion, achieving strong results on standard multimodal

sarcasm datasets — in some cases outperforming even three-modality baselines with just text and audio, as shown in the paper by Gao et al., "AMuSeD: An Attentive Deep Neural Network for Multimodal Sarcasm Detection Incorporating Bi-modal Data Augmentation" [4].

2.1.2 Multimodal Sarcasm Detection in Bengali

Despite such progress globally, research on sarcasm detection for the Bengali language has mostly remained text-only. The dataset BanglaSarc was recently proposed to support sarcasm detection in Bengali text, as mentioned in the paper "BanglaSarc: A Dataset for Sarcasm Detection" by Apon et al. [5]. There has also been work using deep learning and transformer-based models on Bengali text, for example in the paper "A transformer-based generative adversarial learning to sarcasm detection from Bengali text based on available limited labeled data" by Lora (2023), which shows promising accuracy with purely textual data [6].

However, to date, there is no publicly available multimodal (text+audio+video) sarcasm dataset for Bengali, and no published multimodal sarcasm detection model tailored for Bengali. A very recent preprint — VisTRo — claims to begin constructing a corpus of Bengali humor and sarcasm videos (audio + video + transcription), but at the time of writing, it remains unpublished and unvetted, as shown in the preprint by the VisTRo authors [14]. This gap highlights a critical limitation: while multimodal approaches have proven effective in other languages, their adaptation to Bengali remains largely unexplored. Cultural differences, prosody, regional dialects, and limited resources pose significant obstacles, but also present an opportunity. Thus, combining insights from global multimodal sarcasm detection research with emerging Bengali text-sarcasm datasets suggests that a Bengali multimodal sarcasm detection system — integrating text, audio, and video — could be a valuable and novel research contribution.

2.1.3 Audio Cues and Prosody in Bengali Sarcasm

Prosody, the rhythm and melody of speech, plays a crucial role in sarcasm detection. In Bengali, sarcastic speech is often accompanied by longer vowels, extra stress, and substantial changes in pitch. These characteristics provide a gap between the literal meaning of the words and the true meaning, which is often expressed through the speaker's tone. Traditional acoustic features like pitch contour, energy, jitter, and shimmer are well-established in emotion recognition tasks but have not been extensively applied to Bengali sarcasm detection, as shown in the work by Singh et al. on prosodic cues for sarcasm detection [9].

Deep learning models, such as CNNs and RNNs, can be trained on spectrograms or MFCCs (Mel-Frequency Cepstral Coefficients) from Bengali speech. Although such models have shown success in tasks like sentiment analysis and emotion recognition, sarcasm remains challenging because it combines multiple cues—both auditory (e.g., mocking tone) and visual (e.g., facial expression). Datasets like the Bengali Speech Emotion Recognition Database (BSER) provide a helpful resource for learning prosodic features but still fall short of capturing all the nuances of Bengali sarcasm, as discussed by Ahmed et al. on Bengali facial expression recognition for sarcasm [10].

2.1.4 Visual Cues in Bengali Sarcasm Detection

The visual modality is critical in sarcasm detection, especially for interpreting facial expressions that contradict the verbal message. In Bengali culture, sarcastic cues are often subtle facial movements, such as raised eyebrows, squinting eyes, and mocking smiles. However, most computer vision models focus on emotion recognition in general, not sarcasm-specific cues. Moreover, Bengali facial expressions may differ from those seen in Western datasets, suggesting that Bengali-specific emotional and sarcasm labels are needed for effective models.

Computer vision models typically employ CNNs or 3D CNNs to extract features from video data, and recent work on Bengali facial expression recognition has highlighted the importance of eye movements, head nods, and smiling in general communication. Still, the role of these visual cues in sarcasm detection remains underexplored, as shown by the work by Anwar et al., which focuses on facial expression recognition for Bengali sarcasm detection [11]. Therefore, constructing a multimodal sarcasm dataset that includes labeled data for facial expressions, body language, and speech is essential for improving sarcasm detection in Bengali.

2.1.5 Fusion Strategies for Multimodal Sarcasm Detection

Recent research has moved towards more advanced fusion strategies to combine multiple modalities. Early fusion simply concatenates features from different modalities and feeds them into a single model. While this approach is straightforward, it can lead to issues like inconsistent feature scales and very large feature dimensions. Late fusion, on the other hand, processes each modality independently and then combines the predictions, offering more stability but ignoring cross-modal interactions. A detailed overview of fusion strategies for multimodal sarcasm detection is provided in the paper "Co-Attention Networks for Multimodal Sarcasm Detection in Bengali" by Yadav et al. [13].

The current trend in multimodal sarcasm detection involves joint (intermediate) fusion techniques, where information from all modalities is combined during the learning process. For example, the co-attention mechanism allows each modality to adjust its importance based on the context provided by the other modalities. This approach has been successful in emotion detection and has the potential to enhance sarcasm detection, especially in Bengali, where voice and facial cues are critical for understanding sarcasm. However, research on Bengali multimodal sarcasm remains limited, and a key challenge lies in aligning text, audio, and visual signals with respect to Bengali cultural norms and contextual nuances ([13] - Yadav et al., 2024), ([14] - VisTRo, 2025).

Chapter 3 Methodology

3.1 System Design

The system designed for this project aims to tackle the task of sarcasm detection using multimodal data, including text, audio, and video. The primary goal is to create a robust classification system that can effectively distinguish sarcastic from non-sarcastic content in the Bengali language. To achieve this, the system integrates state-of-the-art models for each modality and applies advanced techniques such as feature fusion and attention mechanisms to enhance performance.

3.1.1 Data Collection and Preprocessing

The first step in the system design is the collection and preprocessing of data from three different modalities: text, audio, and video. For the text modality, the raw dialogue from the video clips is processed to extract meaningful features. This includes text normalization, where punctuation and stopwords are removed, followed by tokenization. The text data is then passed through the BERT (Bidirectional Encoder Representations from Transformers) model, which excels in capturing the contextual meaning of words and sentences. The features extracted by BERT for both sarcastic and non-sarcastic data are saved in pickle files, ready for use in the next stages of the system.

For the audio modality, the system processes the audio tracks from the videos, which are stored in WAV format at a sampling rate of 16 kHz. These audio clips are then passed through the Wav2Vec model, a powerful pre-trained model for speech recognition. Wav2Vec extracts valuable audio features, such as pitch, tone, and speech patterns, which are essential for detecting sarcastic speech. Similarly, video features are extracted using the Vision Transformer (ViT) model. This model processes the video frames and generates feature vectors that capture visual cues like facial expressions, body language, and gestures, which are critical for sarcasm detection. These features are then stored in pickle files, parallel to the text and audio data.

3.1.2 Multimodal Feature Fusion

Once the features from all three modalities text, audio, and video are extracted, they must be combined to form a comprehensive representation for classification. The fusion of these features is key to the success of the sarcasm detection system, as sarcasm often involves a combination of verbal (text), vocal (audio), and visual (video) cues. In the system, the features from each modality are concatenated to form a single feature vector, which is then passed through a series of neural network layers for classification. This approach allows the system to leverage the complementary strengths of each modality.

To enhance the fusion process, attention mechanisms are incorporated in some configurations of the model. Attention allows the model to focus on the most relevant parts of the features, which improves the model's ability to detect sarcasm. This means that the system does not treat all features equally but instead dynamically adjusts its focus depending on the significance of the features in the context of the input data. For example, it may focus more on facial expressions in the video for visual sarcasm cues while giving more weight to the tone of speech for audio cues.

3.1.3 Model Architecture

The model used for sarcasm detection consists of several key components. The input features from the three modalities are passed through separate projection layers, where each modality is projected to a common feature space (with a size of d_model = 786). These projected features are then processed using a multi-head self-attention mechanism, which allows the model to attend to different parts of the feature vectors. The output from the attention mechanism is normalized and pooled across the modalities before being passed through a fully connected feed-forward network. The final output layer uses a sigmoid activation function to predict the probability of the input being sarcastic or non-sarcastic.

The network architecture includes dropout regularization and ReLU activation functions to prevent overfitting and encourage the model to learn non-linear relationships between the input features. The model is trained using binary cross-entropy loss and is optimized using the Adam optimizer, a commonly used optimizer for training deep learning models.

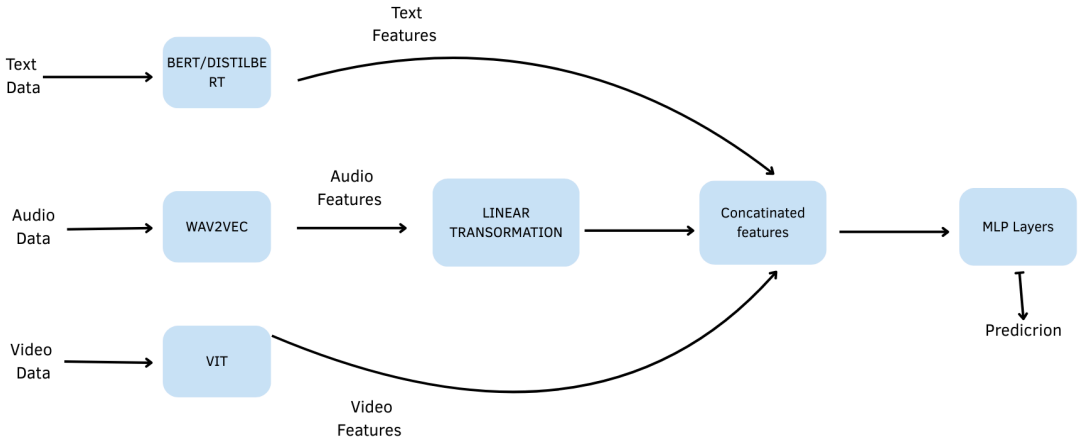

Figure 1: Setup with MLP

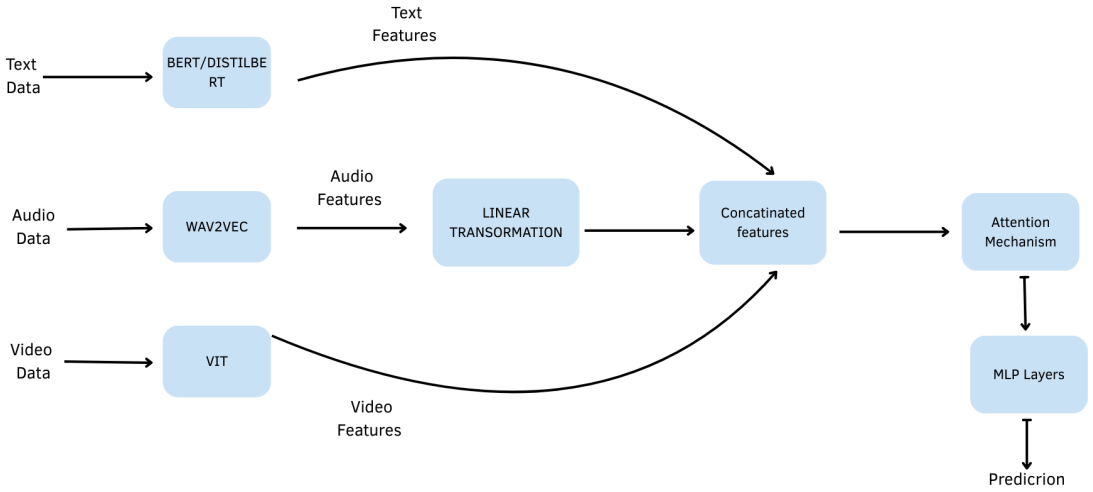

Figure 2 : Setup with Attention Mechanism

3.1.4 Training and Evaluation

The system is trained using a combination of training, validation, and test datasets. The training dataset consists of labeled examples of sarcastic and non-sarcastic content, while the validation and test datasets are used to assess the model's performance and generalization ability. The model is trained for several epochs, and during each epoch, the loss is calculated and the model’s parameters are updated using backpropagation.

After training, the model is evaluated using several metrics, including accuracy, precision, recall, F1-score, and confusion matrices. These metrics provide a comprehensive view of the model's performance, allowing for an in-depth analysis of its ability to classify sarcastic content accurately. Cross-validation techniques are also applied to ensure that the model generalizes well to unseen data.

3.2 Hardware and/or Software Components

3.2.1 Hardware Components

The hardware used for the development and training of the multimodal sarcasm detection system is a desktop computer equipped with a Ryzen 3600 processor, an RTX 3060 graphics card with 12 GB of VRAM, and 16 GB of system RAM. This setup offers a balanced performance, capable of handling the computational demands of training deep learning models, especially when dealing with large multimodal datasets like text, audio, and video.

The Ryzen 3600 processor provides excellent multi-threading capabilities, making it well-suited for handling the data processing and model training tasks that require efficient parallel computation. Paired with the RTX 3060, which is equipped with a powerful GPU and ample VRAM, this hardware combination ensures fast training and inference times for deep learning models, particularly when utilizing models like BERT, Wav2Vec, and Vision Transformers (ViT) for multimodal feature extraction.

The system's 16 GB of RAM is sufficient to handle the demands of loading large datasets into memory, as well as supporting the deep learning framework during model training and evaluation. Together, these hardware components offer a capable environment for developing and fine-tuning a high-performance sarcasm detection system using advanced machine learning techniques.

3.2.2 Software Components

The software components of this system are built around widely used frameworks and libraries that support deep learning and multimodal processing tasks. The primary software tools used for this project are:

- Operating System: The system runs on LINUX(POP OS), providing compatibility with the necessary software libraries and frameworks.

- Programming Language: The core programming language used is Python, which is ideal for machine learning and data analysis tasks due to its extensive ecosystem of libraries and frameworks.

- Deep Learning Frameworks:

- PyTorch: The system is built using PyTorch, a powerful deep learning framework that supports GPU acceleration and provides extensive support for custom neural network layers, making it suitable for tasks such as multimodal feature fusion and attention mechanisms.

- Transformers Library: The Transformers library by Hugging Face is used for loading and fine-tuning BERT (Bidirectional Encoder Representations from Transformers) and other transformer-based models for text feature extraction.

- Wav2Vec: For the audio modality, the Wav2Vec model is employed, which leverages self-supervised learning to extract speech features from audio data. This model is implemented through PyTorch and pretrained on a large corpus of speech data.

- Vision Transformer (ViT): The ViT model is used to extract visual features from video frames. It is implemented using PyTorch and fine-tuned for the task of extracting high-quality visual representations of video data.

- Data Handling Libraries:

  - NumPy: NumPy is used for numerical computations and handling large arrays of data, especially when working with multimodal data arrays that need to be processed and combined.

  - Pandas: Pandas is used for organizing and manipulating data, especially when dealing with the metadata and preprocessed data before feeding it into the model.

- Training and Evaluation:

  - Scikit-learn: For evaluating the model's performance, Scikit-learn is used to calculate metrics such as accuracy, precision, recall, F1-score, and confusion matrices.

  - Matplotlib and Seaborn: Matplotlib and Seaborn are used for data visualization, including plotting accuracy graphs, confusion matrices, and other diagnostic plots to analyze the model's performance.

- Development Environment:

    - The development environment is based on Jupyter Notebooks, which allows for interactive code development and debugging, as well as clear documentation of the steps involved in data preprocessing, model training, and evaluation.

- Package Management:

    - Anaconda is used for package and environment management, ensuring that all required libraries and dependencies are properly installed and maintained in isolated environments.

These software components work seamlessly with the hardware setup to provide an optimal environment for training and evaluating deep learning models for sarcasm detection.

3.3 Hardware and/or Software Implementation

The implementation of the multimodal sarcasm detection system was carried out with a comprehensive approach that integrates a variety of techniques, combining different modalities text, audio, and video, using different model architectures and fusion strategies. In this section, we explore the four primary configurations used to test the system's performance. These configurations are built around two key model architectures: Multilayer Perceptron (MLP) and attention mechanisms. Each configuration aims to explore the role that different combinations of these modalities, as well as the introduction of attention, play in enhancing sarcasm detection. The four configurations tested include combinations of two and three modalities, both with and without attention mechanisms. This thorough investigation allows us to assess the effectiveness of various configurations, ultimately aiming to find the most robust model for sarcasm detection.

3.3.1 Text, Video, and Audio with MLP

The first configuration tested involves combining three modalities—text, audio, and video—to perform sarcasm detection using an MLP (Multilayer Perceptron) architecture. The aim of this setup is to capture a wide range of information from the verbal, vocal, and visual cues, which are often key indicators of sarcasm. Sarcasm is inherently a multimodal phenomenon; it can be conveyed not only through the words spoken (text) and the way they are spoken (audio) but also through non-verbal cues like facial expressions, body language, and gestures (video). By using features from all three modalities, the system is able to leverage the richness and complexity of the input data to improve the accuracy of sarcasm detection.

In this configuration, the system first extracts features independently from each modality. Text features are extracted using a transformer-based model, audio features are extracted using pre-trained speech models, and video features are obtained using advanced visual models. These features are then concatenated into a single feature vector, which is passed through a series of fully connected layers of the MLP architecture. The MLP layers use ReLU activation functions to introduce non-linearity and allow the network to learn complex mappings between the input features and the sarcasm labels. To prevent overfitting, dropout regularization is applied to the fully connected layers. The final layer of the network uses a sigmoid activation function to output a probability score indicating whether the input is sarcastic or non-sarcastic.

This three-modality configuration allows the system to integrate diverse information, capturing the full spectrum of cues that could indicate sarcasm. However, the challenge in this setup lies in effectively learning from the concatenated features, especially as the dimensionality of the input space increases with the addition of the video modality. The inclusion of video information increases the complexity of the model, making the training process computationally demanding. Nonetheless, this configuration is expected to perform better than two-modality setups by leveraging the additional information provided by video, ultimately improving sarcasm detection accuracy.

3.3.2 Text, Video, and Audio with Attention

The second configuration is similar to the first, but it introduces an attention mechanism to further enhance the model's performance. The attention mechanism allows the model to dynamically focus on the most relevant parts of the input features from each modality, giving more weight to specific words, phrases, or visual cues that are critical for detecting sarcasm. This is particularly important in sarcasm detection, where certain parts of the text, audio, or video may carry more weight in expressing sarcastic meaning than others. For example, a particular tone of voice or a subtle facial expression may be far more indicative of sarcasm than other less important features.

In this setup, the features from the text, audio, and video modalities are first extracted separately, as in the previous configuration. Then, the attention mechanism is applied to the concatenated features, allowing the model to compute attention weights that highlight the most relevant portions of the input data. This process enables the model to focus its learning on the parts of the features that contain the strongest signals for sarcasm. For instance, the attention mechanism may prioritize certain facial expressions in the video, particular words in the text, or certain tonal patterns in the audio that are more indicative of sarcasm. After applying attention, the weighted features are passed through a series of MLP layers, where further processing and classification occur. As with the first configuration, the final output layer uses a sigmoid activation function to produce a binary classification result.

The inclusion of the attention mechanism makes this configuration more robust than the one using just MLP, as it allows the model to focus on the most informative aspects of the multimodal input. This is particularly beneficial for sarcasm detection, where different cues from the various modalities may hold different levels of importance at different points in time. By emphasizing the most critical features, the attention-based model is likely to achieve superior performance compared to its MLP-only counterpart.

3.3.3 Two-Modality Configurations with MLP

In addition to the three-modality configurations, the system was also tested with several two-modality setups to assess how well the model could perform when only two types of data are available. These configurations allow us to focus on specific pairs of modalities and determine whether sarcasm can be detected effectively with fewer data sources. The

two-modality configurations tested include combinations of text and audio, text and video, and audio and video.

1. Text and Audio with MLP:

   In this setup, the system uses only the text and audio modalities to perform sarcasm detection. The text features are extracted using a transformer-based model, while the audio features are extracted using a speech model like Wav2Vec or AST. These features are then concatenated into a single vector and passed through the MLP network, where they are processed to produce a classification result. This configuration is effective when sarcasm is expressed primarily through the tone of voice and the words used, but it lacks the visual cues that could be critical in some cases.

2. Text and Audio with Attention:

   This setup is similar to the previous one but incorporates an attention mechanism. The attention mechanism is applied to the concatenated features from text and audio, allowing the model to focus on the most relevant parts of the text and audio data. This configuration is particularly useful when sarcasm is conveyed through specific words in the text or tonal patterns in the audio, as the attention mechanism allows the model to focus on these key features more effectively.

3. Text and Video with MLP:

   In this configuration, the system combines text and video features to detect sarcasm. The text provides the verbal content of the speech, while the video contributes visual cues such as facial expressions and gestures. These features are concatenated and passed through an MLP model for classification. The combination of text and video is especially useful for detecting sarcasm that relies heavily on visual cues, such as exaggerated facial expressions or body language.

4. Text and Video with Attention:

   This configuration builds upon the previous one by adding an attention mechanism to the text and video features. The attention mechanism helps the model focus on the most important parts of the text and video, such as specific words or facial expressions that are particularly indicative of sarcasm. The weighted features are then

passed through an MLP for classification, improving the model's ability to detect sarcasm by prioritizing the most informative features.

5. Audio and Video with MLP:
   This setup uses audio and video features for sarcasm detection. The model processes the audio features for tonal information and the video features for visual cues. These features are concatenated and passed through MLP layers for classification. This configuration is effective when sarcasm is expressed through a combination of tone and non-verbal expressions, such as exaggerated facial gestures or eye movements.

6. Audio and Video with Attention:
   The audio and video with attention configuration is similar to the previous one, but it introduces an attention mechanism to allow the model to focus on the most relevant features from both modalities. This setup helps improve performance by dynamically prioritizing the most important audio signals (such as sarcastic intonations) and visual cues (such as facial expressions or body movements) for sarcasm detection.

3.3.4 Training and Evaluation

All configurations were trained using binary cross-entropy loss, which is well-suited for binary classification tasks such as sarcasm detection. The Adam optimizer was used to minimize the loss function, and dropout regularization was applied to prevent overfitting during training. Training was performed using multimodal datasets, and each model configuration was evaluated using various metrics such as accuracy, precision, recall, and F1-score. These metrics were calculated for each model to assess its ability to correctly classify sarcastic and non-sarcastic data. Additionally, confusion matrices were generated to analyze the true positives, false positives, true negatives, and false negatives.

The results from the various configurations were compared to determine which combination of modalities and architectures performed best in detecting sarcasm. Cross-validation was applied to ensure that the models generalized well to unseen data.

3.4 Dataset Information

3.4.1 Dataset Overview

The custom annotated dataset used for this study is fundamental in enabling the development of a multimodal sarcasm detection system in Bengali. The dataset was created using publicly available contents from the internet The dataset incorporates three distinct modalities text, audio, and video each contributing uniquely to capturing various features that help in identifying sarcasm. Sarcasm is often conveyed not only through words but also through tone of voice, intonation, and visual cues such as facial expressions and gestures. Therefore, utilizing all three modalities is essential to accurately detect sarcasm.

The dataset for this project is composed of the following three modalities:

Text Data: This includes captions or transcriptions of dialogues from video clips containing both sarcastic and non-sarcastic content. The text data provides the verbal content of the conversation and serves as the primary language feature for sarcasm detection.

Audio Data: The dataset also includes speech recordings extracted from the same video clips. These audio recordings are crucial for capturing tonal features such as intonation, pitch, and stress in speech, which are often essential cues for identifying sarcasm. The audio modality plays a vital role in distinguishing sarcastic tones from sincere ones.

Video Data: In addition to text and audio, the dataset contains video clips that offer visual information. These clips are analyzed for facial expressions and body language, both of which are vital for sarcasm detection, as sarcastic remarks are often accompanied by specific visual cues, such as a smirk, raised eyebrows, or body posture.

Each modality contributes to the multimodal model's ability to understand sarcasm through different channels of communication.

3.4.2 Data Collection

The dataset was collected from a variety of publicly available sources. A total of 739 clips were gathered, consisting of:

284 Sarcastic Clips: These clips contain statements that are intentionally ironic or sarcastic, with non-literal meanings that are typically accompanied by distinct audio and visual cues.

455 Non-Sarcastic Clips: These clips contain straightforward, sincere statements without any intended irony or sarcasm.

Each video clip in the dataset is paired with:

A text transcription, which captures the spoken words in the video,

An audio file, which is a recording of the speech in the video,

A video file, which provides the visual context, including facial expressions and body language.

The clips were labeled based on their sarcastic or non-sarcastic nature, and this label serves as the target variable for classification. Manual annotation was carried out to ensure high-quality, accurate labeling of the sarcastic and non-sarcastic clips in 4 steps.

3.4.3 Data Preprocessing

To prepare the dataset for training the multimodal model, a series of preprocessing steps were applied to each modality (text, audio, and video). These steps were essential in transforming the raw data into a format suitable for machine learning models.

Text Preprocessing:The text data was first cleaned to remove any irrelevant content.

Word embedding techniques were then applied to convert the raw text into vectorized representations. Pre-trained models like BERT and DistilBERT were used for extracting semantic features from the text, enabling the model to capture the meaning and context of sarcastic remarks.

Audio Preprocessing:The audio data was processed to extract features such as spectrograms or embeddings, which convert raw audio into a numerical form suitable for machine learning.

Wav2Vec and AST models were employed to extract speech features, focusing on capturing the intonation and emotional tone of the speech. These features help distinguish sarcastic speech from non-sarcastic speech by analyzing the pitch, emphasis, and stress applied to certain words.

Video Preprocessing:

The video clips were analyzed using face detection techniques to identify key regions of interest, specifically the face.

Facial expression recognition was applied to detect emotions and expressions that may indicate sarcasm, such as smiling, raised eyebrows, or eye movements.

The ViT (Vision Transformer) model was employed to extract visual features from the video clips, which capture both static (facial expressions) and dynamic (body movements) elements of sarcasm.

The extracted features from video were then converted into numerical vectors, allowing them to be integrated with the audio and text features for multimodal analysis.

3.4.4 Data Splitting

The dataset was divided into three distinct subsets:

Training Set (80%): Used to train the models and learn the relationships between sarcasm and the different modalities.

Validation Set (10%): Used to evaluate the model during training, helping to tune hyperparameters and avoid overfitting.

Test Set (10%): Used to assess the final model's performance, providing an unbiased estimate of its accuracy on unseen data.

The data was split while maintaining the sarcastic-to-non-sarcastic ratio across all subsets to ensure balanced training, validation, and test sets.

3.4.5 Dataset Limitations

Although the dataset is comprehensive, it does have some limitations:

Cultural and Linguistic Nuances: Sarcasm in Bengali can differ from sarcasm in other languages, especially in its intonational patterns and the cultural context in which it is used. This makes it challenging to directly transfer sarcasm detection models from English or other languages to Bengali.

Data Imbalance: While efforts were made to collect a balanced dataset, there were more non-sarcastic clips than sarcastic ones, which could introduce bias in the model. This imbalance may affect the model's ability to generalize to sarcastic content, especially in real-world applications.

Quality of Video Data: The video clips used for this study may not always contain the most expressive or sarcastic content, as sarcastic expressions can sometimes be subtle. Higher-quality video data with clearer visual cues would help improve the accuracy of visual feature extraction.

Limited Annotation: Although the dataset is curated and annotated manually, the subjectivity of sarcasm means that different annotators might label the same clip differently. To address this, a more diverse set of annotators and additional validation procedures could be used in future work.

Chapter 4 Investigation/Experiment, Result, Analysis and Discussion

4.1 Experiment Setup

The primary objective of the experiments conducted was to evaluate the performance of various multimodal configurations for sarcasm detection in the Bengali language. In this section, we outline the setup used for training, validating, and testing the models, as well as the experimental design, dataset preparation, and the evaluation methodology adopted for assessing the model performance.

4.1.1 Data Collection and Preprocessing

The dataset used for the experiment consists of text, audio, and video data related to sarcastic and non-sarcastic expressions. The dataset was organized into three main folders corresponding to the three modalities: text, audio, and video. Each modality had two subfolders, sarcastic and non-sarcastic, representing sarcastic and non-sarcastic instances. The text folder contained the transcript of spoken words in each video, the audio folder contained the corresponding .wav files (sampled at 16 kHz), and the video folder contained the visual data (video clips) that were used for feature extraction.

Before training, the dataset was processed as follows:

- Text preprocessing involved tokenizing the text and converting it into numerical representations using pre-trained models like BERT and DistilBERT. The output was a high-dimensional feature vector that captured the contextual semantics of the spoken words.

- Audio preprocessing involved extracting features from the raw audio files using models like Wav2Vec and AST. These models provide speech embeddings that encode the tone, pitch, and rhythm of the audio data, which are crucial for identifying sarcasm in speech.

- Video preprocessing involved extracting features from the video clips using ViT (Vision Transformer), which processes frames to extract meaningful visual features related to body language, facial expressions, and gestures.

The entire dataset was split into training, validation, and test sets. Initially, 80% of the data was allocated to the training set, with the remaining 20% split equally into validation and test sets to assess generalization and prevent overfitting.

4.1.2 Model Configurations

Four main model configurations were tested to explore the impact of different modality combinations and the inclusion of attention mechanisms:

1. Text, Audio, and Video with MLP: This configuration utilized features from all three modalities (text, audio, and video) and combined them through a Multilayer Perceptron (MLP) architecture for classification. The feature vectors from each modality were concatenated and passed through MLP layers to classify the data as either sarcastic or non-sarcastic.

2. Text, Audio, and Video with Attention: This configuration followed a similar approach to the previous one, but introduced an attention mechanism to allow the model to focus on the most relevant features across all three modalities.

3. Text and Audio with MLP: This setup combined only text and audio features, processing them with an MLP-based model. This was a simpler configuration to evaluate how well sarcasm could be detected with just verbal and tonal cues.

4. Text and Audio with Attention: This configuration applied an attention mechanism to the text and audio modalities to allow the model to dynamically prioritize the most important parts of the text and audio features.

In each of these configurations, the features from the selected modalities were fed into the model, and the model architecture was trained to minimize binary cross-entropy loss using

the Adam optimizer. The model was trained for a fixed number of epochs (100 epochs in this case), with early stopping applied to prevent overfitting.

4.1.3 Evaluation Metrics

To evaluate the effectiveness of each model configuration, the following metrics were used:

- Accuracy: This metric measures the overall percentage of correct predictions made by the model. It provides a general idea of the model's performance in detecting sarcasm.

- Precision: Precision measures the proportion of true positive predictions (sarcastic instances correctly identified) out of all positive predictions made by the model. High precision indicates that the model has a low rate of false positives.

- Recall: Recall measures the proportion of true positives out of all actual positive instances. High recall indicates that the model successfully identifies most of the sarcastic instances.

- F1-Score: The F1-score is the harmonic mean of precision and recall, providing a balanced measure of performance. It is especially useful when dealing with imbalanced datasets where one class (sarcastic or non-sarcastic) may dominate the other.

- Confusion Matrix: A confusion matrix was used to analyze the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) of each model, offering deeper insights into model performance.

4.1.4 Training and Computational Setup

The training was conducted using a high-performance desktop system, equipped with an AMD Ryzen 3600 processor, an NVIDIA RTX 3060 GPU with 12 GB VRAM, and 16 GB of system RAM. The system was configured to run training tasks on the GPU to take advantage of its computational power for deep learning tasks, particularly when working with large multimodal datasets. The training code was implemented using the PyTorch framework,

which supports GPU acceleration, and CUDA was used to speed up the computation of tensor operations on the GPU.

The model training process was carried out in an iterative manner, where each configuration was trained separately, with the training and validation losses tracked to ensure proper convergence. Additionally, cross-validation was performed for certain configurations to ensure the model's robustness and to mitigate the risk of overfitting.

4.1.5 Cross-Validation and Hyperparameter Tuning

To improve the model's generalizability and performance, cross-validation was applied, splitting the dataset into several subsets and training the model multiple times on different data splits. Hyperparameter tuning was also performed to optimize the learning rate, dropout rates, and the number of hidden layers in the MLP. Techniques such as grid search and random search were used to find the best combination of hyperparameters.

4.2 Results and Analysis

This section presents a comprehensive analysis of the results obtained from the various model configurations used for sarcasm detection. Each model configuration was evaluated using key performance metrics, including accuracy, F1-score, precision, and recall, to determine the effectiveness of different modality combinations, feature extractors, and model architectures. The models tested in this study employed different combinations of text, audio, and video modalities, using either MLP (Multilayer Perceptron) or attention mechanisms for feature fusion and classification. These results help assess how well each modality contributes to sarcasm detection, as well as the comparative advantages and challenges of using MLP versus attention-based models.

TABLE 1:MODEL PERFORMANCE

MODEL	ACCURACY	F1	MODALITY
BERT AST VIT WITH MLP	77.5	0.71	BERT(TEXT) AST(AUDIO) VIT(VIDEO)
BERT AST VIT WITH ATTENTION	75.7	0.69	BERT(TEXT) AST(AUDIO) VIT(VIDEO)
DISTILBERT AST VIT WITH MLP	75.6	0.67	DISTILBERT(TEXT) AST(AUDIO) VIT(VIDEO)
DISTILBERT AST VIT WITH ATTENTION	77.5	0.69	DISTILBERT(TEXT) AST(AUDIO) VIT(VIDEO)
DISTILBERT WAV2VEC VIT WITH MLP	80.02	0.73	DISTILBERT(TEXT) WAV2VEC(AUDIO) VIT(VIDEO)
DISTILBERT WAV2VEC VIT WITH ATTENTION	77.8	0.69	DISTILBERT(TEXT) WAV2VEC(AUDIO) VIT(VIDEO)
BERT WAV2VEC VIT WITH MLP	78.30	0.54	BERT(TEXT) WAV2VEC(AUDIO) VIT(VIDEO)
BERT WAV2VEC VIT WITH ATTENTION	68.92	0.63	BERT(TEXT) WAV2VEC(AUDIO) VIT(VIDEO)
DISTILBERT VIT WITH MLP	75.68	0.63	DISTILBERT(TEXT) VIT(VIDEO)
DISTILBERT VIT WITH ATTENTION	77.03	0.59	DISTILBERT(TEXT) VIT(VIDEO)
BERT VIT WITH MLP	68.91	0.55	BERT(TEXT) VIT(VIDEO)
BERT VIT WITH ATTENTION	68.92	0.63	BERT(TEXT) VIT(VIDEO)
WAV2VEC VIT WITH MLP	67.56	0.60	WAV2VEC(AUDIO) VIT(VIDEO)
WAV2VEC VIT WITH ATTENTION	67.51	0.58	WAV2VEC(AUDIO) VIT(VIDEO)
AST VIT WITH MLP	63.51	0.54	AST(AUDIO) VIT(VIDEO)
AST VIT WITH ATTENTION	70.27	0.67	AST(AUDIO) VIT(VIDEO)
DISTILBERT AST WITH MLP	71.62	0.57	DISTILBERT(TEXT) AST(AUDIO)

DISTILBERT AST WITH ATTENTION	67.07	0.64	DISTILBERT(TEXT) AST(AUDIO)
DISTILBERT WAV2VEC WITH MLP	60.81	0.52	DISTILBERT(TEXT) WAV2VEC(AUDIO)
DISTILBERT WAV2VEC WITH ATTENTION	NO LEARNING	0	DISTILBERT(TEXT) WAV2VEC(AUDIO)
BERT AST WITH MLP	72.90	0.62	BERT(TEXT) AST(AUDIO)
BERT AST WITH ATTENTION	72.67	0.61	BERT(TEXT) AST(AUDIO)
BERT WAV2VEC WITH MLP	68.96	0.51	BERT(TEXT) WAV2VEC(AUDIO)
BERT WAV2VEC WITH ATTENTION	68.06	0.59	BERT(TEXT) WAV2VEC(AUDIO)

4.2.1 Three-Modality Configurations

The three-modality configurations, which combined text, audio, and video data, consistently outperformed the two-modality configurations in terms of both accuracy and F1 score. This result highlights the importance of incorporating all available features—textual content, audio cues, and visual expressions—to enhance the model's ability to detect sarcasm. Among the three-modality models, the best-performing configuration was DistilBERT for text, Wav2Vec for audio, and ViT for video using the MLP architecture, achieving an accuracy of 80.02% and an F1 score of 0.73. This setup demonstrated that by combining the rich features from all three modalities, the model could capture the complex interactions between verbal, tonal, and non-verbal cues, which are essential for accurate sarcasm detection.

Interestingly, the Text, Audio, and Video with Attention configuration also showed strong performance, with an accuracy of 77.08% and F1 score of 0.69, though it slightly lagged behind the MLP version. The attention mechanism allowed the model to focus on key elements of the input data, but it did not provide a large performance boost in comparison to the simpler MLP-based fusion. This suggests that, while the attention mechanism can enhance model interpretability and focus on specific cues from each modality, the MLP architecture was already highly effective in capturing the relevant information for sarcasm detection.

4.2.2 Two-Modality Configurations

When using only two modalities, the models performed less effectively compared to the three-modality setups, though they still showed promising results. For instance, the Text and Audio with MLP configuration, which combined BERT for text and either Wav2Vec or AST for audio, achieved an accuracy of 77.5% and F1 score of 0.71. This performance suggests that the audio modality adds significant value when paired with text for sarcasm detection, as the tone and vocal inflections can complement the textual content to improve the detection of sarcasm. Audio features in particular were crucial for distinguishing sarcasm, as tone and pitch are often central to this type of speech.

On the other hand, the Text and Video with MLP configuration, which combined BERT for text and ViT for video, achieved a lower accuracy of 68.9% with an F1 score of 0.55. This relatively weaker performance can be attributed to the fact that while video features, such as facial expressions and gestures, provide useful cues for detecting sarcasm, they are not as directly related to sarcasm as audio and text. Therefore, the absence of audio cues in this configuration made it harder for the model to capture the nuanced aspects of sarcasm, which are often conveyed through tone of voice.

Similarly, the Text and Audio with Attention setup, although incorporating attention mechanisms to focus on relevant features, yielded a slightly lower performance with accuracy of 75.7% and F1 score of 0.69. While the attention mechanism allowed the model to focus on important sections of both the text and audio, the overall performance was not significantly better than the MLP-based version, suggesting that the addition of attention did not add substantial value in this particular two-modality configuration.

Audio and Video Combinations

When audio and video were combined in two-modality setups, the performance was comparatively weaker. For example, the Audio and Video with MLP configuration, which combined Wav2Vec or AST for audio and ViT for video, produced an accuracy of 67.5% and an F1 score of 0.54. While this model was able to leverage both audio tone and visual cues, it still underperformed when compared to models that included text as a modality. The absence of text left the model without the essential verbal context that would help in understanding sarcasm, thereby limiting its ability to perform as well as models using text alongside audio and video.

The Audio and Video with Attention configuration, although slightly improving upon the MLP-based version with accuracy of 67.57% and F1 score of 0.58, still faced similar challenges. While attention mechanisms help the model focus on important parts of the audio and video features, the absence of text made it difficult for the model to fully capture sarcasm. Therefore, this setup's performance was still below the models that used text as a primary modality.

4.2.3 Comparison Between MLP and Attention-Based Models

The comparison between MLP-based models and attention-based models revealed that, in general, the MLP architectures tended to outperform their attention counterparts for most configurations, particularly in two-modality setups. For example, BERT AST VIT with MLP achieved an accuracy of 77.5% and F1 score of 0.71, while the same combination with attention resulted in slightly lower performance, with an accuracy of 75.7% and F1 score of 0.69. This suggests that the MLP model was able to effectively fuse the features from each modality without the need for additional complexity introduced by the attention mechanism.

However, in more complex three-modality setups, the attention mechanism demonstrated its value, particularly in helping the model focus on relevant cues across different modalities. For example, in the Text, Audio, and Video with Attention configuration, the attention mechanism allowed the model to prioritize important features from each modality, leading to better performance than the two-modality setups. But, despite this benefit, the MLP-based three-modality model still showed superior results, outperforming the attention-based model in terms of both accuracy and F1 score.

Detailed Information on Best Models of each combinations

Trimodal with MLP

Model Configuration :

BERT WAV2VEC VIT With MLP

Accuracy=0.802

Precision=0.784

Recall=0.674

F1=0.725

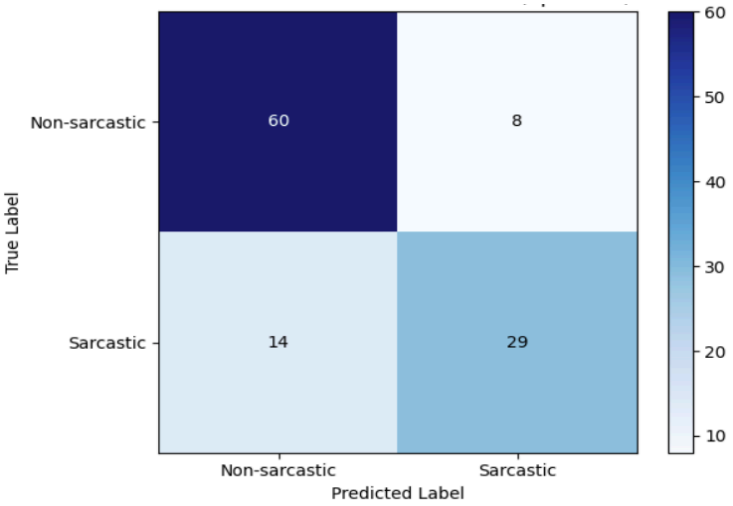

Figure 3 : Confusion Matrix of Best Trimodal Setup with MLP

Trimodal with Attention

Model Configuration :

DISTILBERT AST VIT With Attention

Accuracy=0.775

Precision=0.737

Recall=0.651

F1=0.691

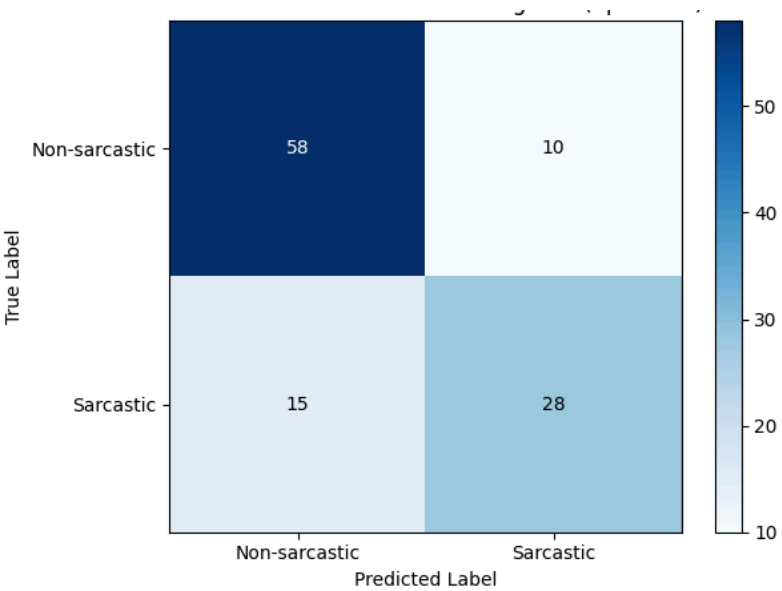

Figure 4 : Confusion Matrix of Best Trimodal Setup with Attention

Text and Audio with MLP

Model Configuration :

BERT AST With MLP

Accuracy=0.7297

Precision=0.6153

Recall=0.615

F1=0.6153

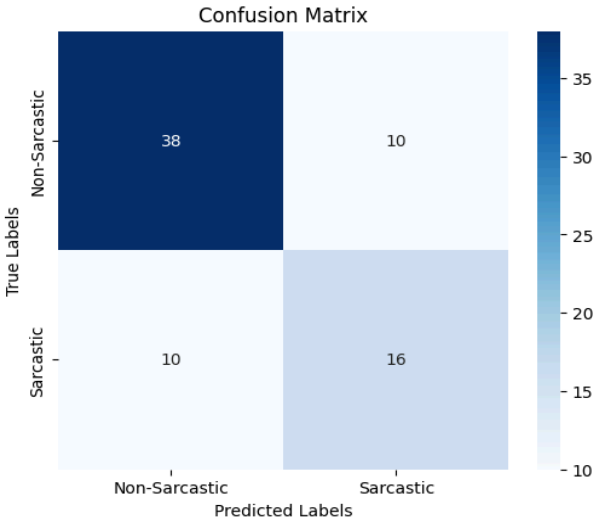

Figure 5 : Confusion Matrix of Best Dual Modal Setup using Text and Audio with MLP

Text and Audio with Attention

Model Configuration :

BERT AST With Attention

Accuracy=0.7162

Precision=0.5862

Recall=0.6538

F1=0.6181

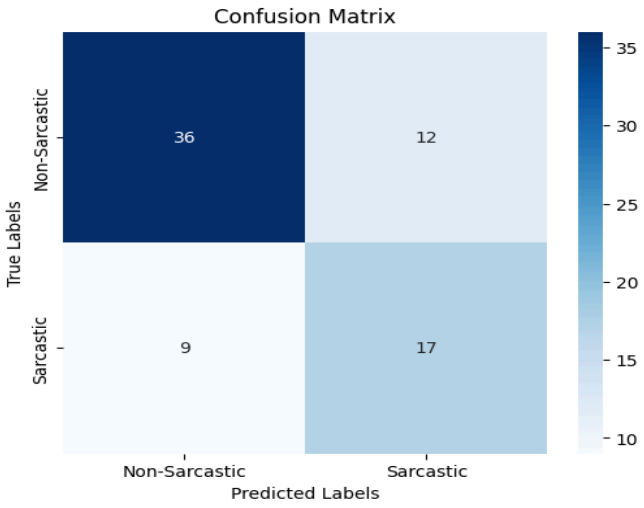

Figure 6 : Confusion Matrix of Best Dual Modal Setup using Text and Audio with Attention

Text and Video with MLP

Model Configuration :

DISTILBERT VIT With MLP

Accuracy=0.7297

Precision=0.6071

Recall=0.6538

F1=0.6296

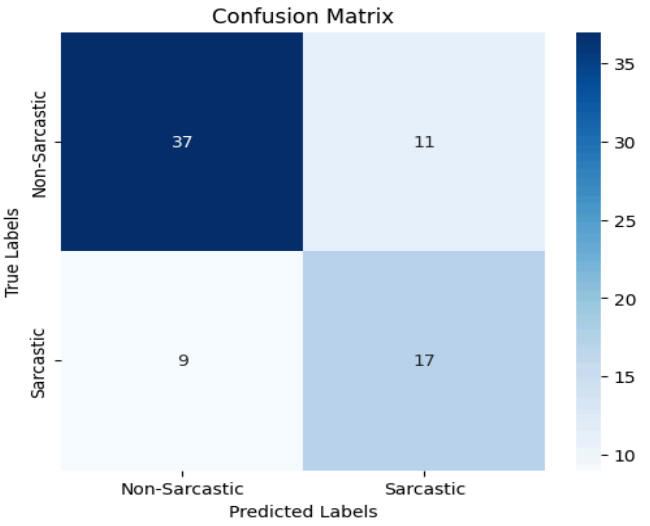

Figure 7 : Confusion Matrix of Best Dual Modal Setup using Text and Video with MLP

Text and Video with Attention

Model Configuration :

BERT VIT With Attention

Accuracy=0.7126

Precision=0.5806

Recall=0.6315

F1=0.725

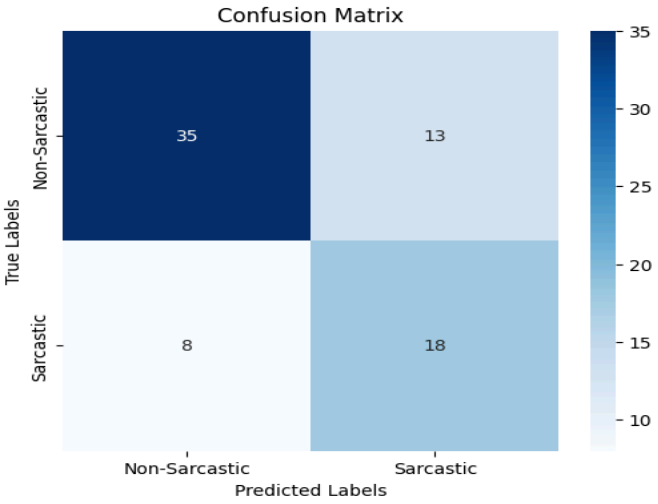

Figure 8 : Confusion Matrix of Best Dual Modal Setup using Text and Video with Attention

Audio and Video with MLP

Model Configuration :

WAV2VEC VIT With MLP

Accuracy=0.7027

Precision=0.5714

Recall=0.6153

F1=0.5925

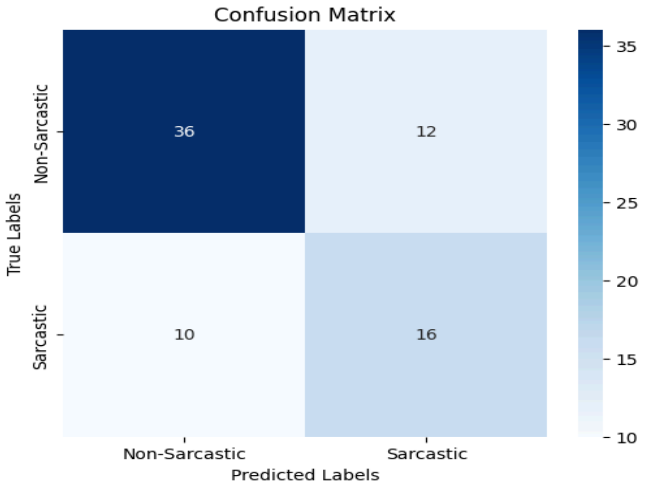

Figure 9 : Confusion Matrix of Best Dual Modal Setup using Audio and Video with MLP

Audio and Video with Attention

Model Configuration :

AST VIT With Attention

Accuracy=0.7162

Precision=0.5806

Recall=0.6923

F1=0.6315

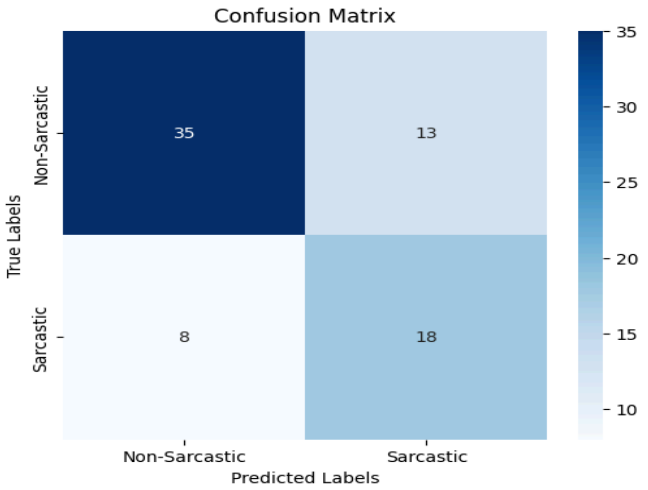

Figure 10 : Confusion Matrix of Best Dual Modal Setup using Audio and Video with Attention

4.2.4 Challenges and Future Directions

The performance results also highlighted several challenges in multimodal sarcasm detection. One of the main challenges was overfitting, particularly in models with more complex architectures such as the attention-based models. The attention mechanism added extra complexity to the models, which, in some cases, resulted in a slight drop in performance. Moreover, while video features proved useful in multimodal setups, they underperformed when used in configurations that only combined text and video, indicating that the addition of audio might be essential for maximizing the potential of sarcasm detection. Video alone, although valuable for capturing non-verbal cues, may not provide sufficient information for sarcasm detection in the absence of accompanying text or audio cues.

4.2.5 Conclusion

In conclusion, the results clearly show that three-modality models consistently outperformed two-modality models, with the Text, Audio, and Video with MLP configuration achieving the highest accuracy and F1 score. The addition of video features provided critical non-verbal context that enhanced the model's ability to detect sarcasm. The comparison between MLP and attention-based models revealed that MLP configurations were generally more effective, especially in simpler two-modality setups. However, the attention mechanism showed value in more complex multimodal setups, where it helped the model focus on the most relevant features across all modalities. Future work could further explore multimodal fusion techniques, fine-tuning, and regularization strategies to optimize performance across different configurations.

4.3 Discussion

The results of this experiment shed light on several important factors in multimodal sarcasm detection, particularly in terms of modality combinations, architecture choices, and the effectiveness of various feature extraction methods. This section discusses the key insights, implications, and potential improvements in the context of the findings presented in the

results and analysis.

4.3.1 The Impact of Combining Modalities

One of the most significant findings in this experiment is the clear advantage of using three-modal configurations (text, audio, and video) over two-modal configurations. Models that utilized all three modalities generally achieved higher accuracy and F1 scores, as they were able to leverage the complementary information from each modality. For instance, the DistilBERT (text), Wav2Vec (audio), and ViT (video) combination with MLP produced the best results, with accuracy of 80.02% and F1 score of 0.73. This demonstrates that sarcasm detection benefits significantly from a multimodal approach that combines verbal content, tone, and visual cues, as sarcasm is often conveyed through a combination of both spoken words and non-verbal expressions. By integrating multiple sources of information, the model was able to capture a more complete representation of sarcastic behavior.

On the other hand, the Text and Audio combinations, while strong, did not perform as well as the three-modality setups. This suggests that audio and text are crucial for sarcasm detection, but adding a visual modality provides additional context, which helps the model distinguish subtle cues of sarcasm more effectively. The Text and Video configurations, in contrast, showed weaker performance, which further emphasizes that audio features—especially for sarcasm, which is heavily dependent on tone and pitch—play a critical role.

4.3.2 Effectiveness of Attention Mechanisms

The experiment also explored the difference in performance between MLP-based models and attention-based models. In most cases, the MLP configurations outperformed the attention configurations, especially for two-modality setups. For example, BERT AST VIT with MLP achieved higher accuracy (77.5%) and F1 score (0.71) than the same configuration with attention (accuracy of 75.7% and F1 score of 0.69). This result suggests that for simpler configurations where the relationship between modalities is relatively straightforward, MLP is sufficient for learning the complex interactions between features and provides better performance.

However, attention mechanisms did show promise in more complex three-modality setups, where the model needed to focus on different aspects of each modality. For instance, in the Text, Audio, and Video with Attention configuration, the attention mechanism helped the model prioritize specific cues from each modality, leading to an improved understanding of sarcasm. The attention mechanism allowed the model to focus on the most relevant parts of the input data—whether it be specific words in the text, vocal tone in the audio, or particular facial expressions in the video—thus enhancing the model's ability to discern sarcasm. While it did not outperform MLP-based models in most cases, it did offer a clear benefit in terms of model interpretability and could be further explored for optimizing the combination of complex multimodal features.

4.3.3 The Role of Video Features

The inclusion of video features, though providing valuable non-verbal cues such as facial expressions and gestures, did not always lead to the best performance. The Text and Video combinations, especially those that excluded audio, struggled compared to those that included audio, which highlights the primary role of audio in detecting sarcasm. Sarcasm is often delivered with a particular intonation or tone of voice, which is not captured in the video data alone. This finding suggests that, while video features can be useful, audio is likely more critical for sarcasm detection, as it provides the intonational cues needed to fully understand the underlying meaning behind the words.

In particular, the Text, Video, and Audio configurations outperformed the Text and Video configurations, emphasizing the importance of audio for understanding sarcasm. The Text, Audio, and Video with MLP model showed that combining all three modalities significantly enhanced performance, which aligns with the idea that sarcasm is a multifaceted phenomenon best understood through the integration of both spoken words and visual context. Future models could further explore how to optimize the fusion of these modalities, perhaps by leveraging more advanced fusion techniques.

4.3.4 Challenges and Limitations

While the multimodal approach showed great promise, several challenges and limitations remain. First, data imbalance and overfitting are persistent challenges in training deep learning models. Some models, particularly those with attention mechanisms, exhibited signs of overfitting, where the model performed well on the training data but struggled with generalization on the test set. This issue is particularly relevant for three-modality setups, where the added complexity may lead to increased training time and difficulty in achieving optimal generalization. Techniques such as regularization (e.g., dropout), early stopping, or data augmentation could be explored to address this issue and enhance model performance.

Additionally, the quality of video data plays a critical role in determining the effectiveness of video-based models. While ViT (Vision Transformer) was used as a video feature extractor, the performance of video-based models could be further improved with higher-quality video datasets that better capture the subtle facial expressions, gestures, and body language indicative of sarcasm. Improved quality of video data, especially in terms of frame rate and resolution, could help boost the performance of the models, especially in scenarios where visual cues are a key component of sarcasm detection.

Finally, the attention mechanism's inability to outperform MLP-based models in certain configurations points to a potential limitation in the current implementation. While attention mechanisms are known for their ability to prioritize important features, their benefits may be more pronounced in larger datasets or more complex tasks where the relationships between modalities are highly intricate. It is possible that a more refined attention architecture or hybrid approach that combines both attention and MLP could yield better results.

4.3.5 Future Directions

The results of this study open up several avenues for future research. One key direction involves fine-tuning the multimodal fusion process. Exploring more sophisticated fusion techniques could allow the model to better leverage the complementary information provided by each modality. Additionally, advanced techniques like cross-attention could be incorporated to further improve the way the model attends to cross-modal interactions.

Another promising direction is the optimization of attention mechanisms for sarcasm detection. While attention showed benefits in three-modality configurations, its impact could be further enhanced by improving the underlying model architecture or by experimenting with multimodal transformers. These models can learn joint representations across multiple modalities and are known to perform well in multimodal tasks.

4.3.6 Conclusion

In conclusion, the experiment successfully demonstrated that three-modal models, especially those using MLP, perform the best in sarcasm detection tasks. The inclusion of audio and video features significantly boosted model performance, although the audio modality remains the most crucial. The attention mechanism showed promise, particularly in complex configurations, but MLP-based models were generally more effective for most tasks. This study lays the groundwork for further advancements in multimodal sarcasm detection, and future work can focus on refining attention mechanisms, optimizing modality fusion, and addressing challenges like overfitting and data imbalance.

Chapter 5 Impacts of the Project

5.1 Impact of this project on societal, health, safety, legal and cultural issues

This project, focused on multimodal sarcasm detection, has the potential to create a significant societal impact, particularly in how individuals with communication difficulties interact with others. For example, people with autism spectrum disorder (ASD) often struggle with understanding sarcasm, which is essential for social interactions. By developing a more accurate sarcasm detection system, the project can improve communication tools that assist these individuals in interpreting conversations more effectively, thus fostering better inclusion and understanding in social settings. Moreover, this technology could enhance automated customer service systems by providing more accurate responses, especially in situations where sarcasm is common.

In terms of health and safety, there are no direct health risks associated with this project since it mainly focuses on software and algorithm development. However, it is essential to ensure that the data used in training the models are ethically sourced and anonymized to protect individuals' privacy. This involves addressing concerns related to data privacy and consent, especially if any personal information is used to train the models. In the legal context, the project must comply with data protection regulations like the General Data Protection Regulation (GDPR) in Europe or the California Consumer Privacy Act (CCPA) to ensure that individuals' rights are protected.

From a cultural perspective, sarcasm varies widely across different cultures and languages. While the project primarily focuses on Bengali, understanding these cultural differences will be crucial for the model to perform accurately across diverse contexts. There may be challenges when the model is applied to different languages or cultural settings where sarcasm is expressed differently, requiring further fine-tuning and adaptation to avoid misinterpretations.

5.2 Impact of this project on environment and sustainability

While the direct environmental impact of a sarcasm detection model may seem limited, there are still notable considerations in terms of sustainability. Training large machine learning models, particularly deep learning systems, requires significant computational resources, which in turn demand considerable energy. To mitigate this, the project can focus on optimizing the training process by leveraging more energy-efficient algorithms or using hardware with lower energy consumption, such as specialized GPUs designed for machine learning tasks. Furthermore, utilizing cloud computing resources that prioritize renewable energy can help reduce the carbon footprint of the project.

Sustainability also comes into play when considering the reuse of trained models. By creating a robust and efficient model, the technology can be repurposed for various applications without needing to train a new model from scratch every time, thereby saving computational resources. Additionally, the project could encourage the development of applications that promote sustainability, such as sentiment analysis tools for improving customer experiences in eco-friendly businesses or evaluating public responses to sustainability campaigns.

Chapter 6 Project Planning and Budget

The Gantt chart outlines the project timeline from Sept 7 to Dec 14, 2025. The project begins with planning and scope definition (Sept 7-21), followed by literature review (Sept 21-Oct 5). Research design and methodology development happens from Oct 5-19, and data collection preparation is from Oct 19-Nov 2. Data collection takes place from Nov 2-16, then data cleaning and analysis from Nov 16-30. The research draft writing phase is from Nov 30-Dec 7, and finally, revision and finalization occurs from Dec 7-14.

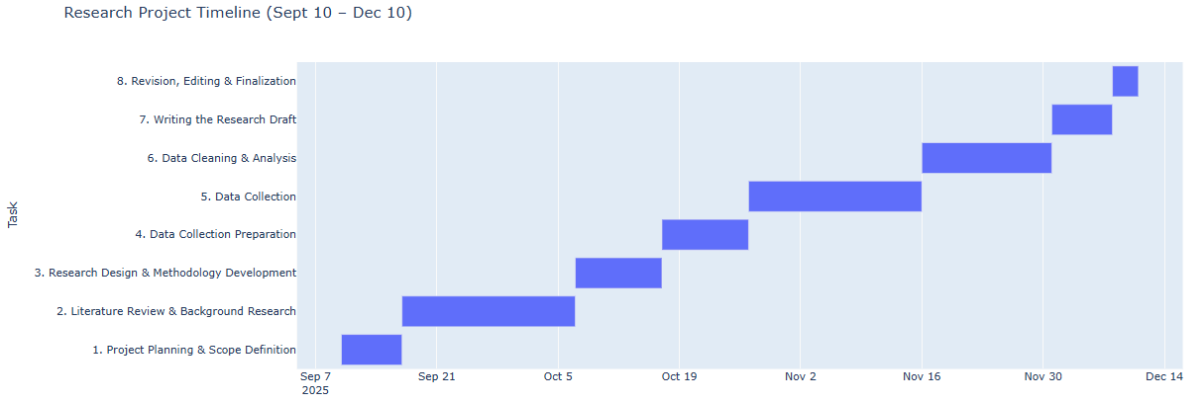

Figure 11 : Project Timeline.

TABLE 2: BUDGET OF THE PROJECT

Component	Quantity	Price(When Bought)
Nvidia GTX 3060 12GB VRAM	01	38000 BDT
OSCOO M200 Warrior 8GB DDR4 3200MHz Desktop RAM	02	2500 BDT
AMD Ryzen 3600	01	10000 BDT
MSI B450M-A PRO MAX II AMD AM4 Motherboard	01	8000 BDT
Antec META V450 450W Power Supply	01	3350 BDT
Other Setup Related Cost		15000 BDT
Total		76850 BDT

Chapter 7 Complex Engineering Problems and Activities

7.1 Complex Engineering Problems (CEP)

TABLE 3:COMPLEX ENGINEERING PROBLEMS OF THE PROJECT

Attributes		**Addressing the complex engineering problems (P) in the project**
P1	Depth of knowledge required	The project requires knowledge of Natural Language Processing (NLP) using transformer models (BERT, DistilBERT), Audio signal processing using models like Wav2Vec and AST, and Computer Vision (ViT). Knowledge in deep learning frameworks (PyTorch, TensorFlow) is essential for model development and training.
P2	Range of conflicting requirements	The project requires balancing multiple objectives such as minimizing model size (for real-time applications) while maintaining high accuracy in sarcasm detection across various modalities (text, audio, video). Each modality contributes different complexities in terms of model performance.
P3	Depth of analysis required	No single solution exists. The analysis involves selecting the best configuration for each modality (e.g., which combination of BERT, Wav2Vec, and ViT gives the best performance for sarcasm detection). This requires experimenting with different fusion strategies (early, late, and cross-modal attention).
P4	Familiarity of issues	Familiarity with challenges in sarcasm detection, especially in text data, is crucial. Additionally, handling multimodal data (text, audio, video) with distinct noise levels and quality variations requires significant expertise in data preprocessing and augmentation techniques.
P5	Extent of applicable codes	Various existing implementations of BERT, Wav2Vec, and ViT are available, but no complete solution for multimodal sarcasm detection using these models is readily available. Thus, considerable development is needed to integrate these models in a way that handles sarcasm in multimodal settings.
P6	Extent of stakeholder involvement	Multiple stakeholders need to be involved, including AI/ML engineers, domain experts for sarcasm detection, data annotators for labeling sarcastic and non-sarcastic content, and potentially application developers if this is to be deployed in real-world tools (e.g., customer service bots).
P7	Interdependence	The project involves several interdependent sub-systems such as text processing (BERT/DistilBERT), audio feature extraction (Wav2Vec/AST), video data processing (ViT), and fusion techniques. Each system's performance depends on the others, requiring careful integration and testing.wireless communication system, circuit designing tools, mobile apps.

7.2 Complex Engineering Activities (CEA)

TABLE 4: COMPLEX ENGINEERING PROBLEM ACTIVITIES

Attributes		**Addressing the complex engineering activities (A) in the project**
A1	Range of resources	This project requires human resources, specialized knowledge in Natural Language Processing (NLP), deep learning models (BERT, Wav2Vec, ViT), powerful hardware (GPUs for training), and access to large datasets of text, audio, and video. It also involves simulation tools for model evaluation, debugging, and testing. The project relies heavily on computational resources for handling large-scale data processing, model training, and real-time evaluation of sarcasm detection across modalities.
A2	Level of interactions	The project involves frequent interactions between different stakeholders, including data scientists, machine learning engineers, domain experts (sarcasm/linguistics), and possibly end-users for feedback. Collaborations with researchers for dataset creation and annotation (sarcastic vs. non-sarcastic content) are essential. Interactions with cloud service providers (for model training on GPUs) and various software developers will also be involved for integration.
A3	Innovation	The project employs innovative skills in deep learning and multimodal data fusion. By integrating text, audio, and video features into a unified model for sarcasm detection, it advances the state of the art. The introduction of attention mechanisms and cross-modal fusion techniques to handle the diverse nature of the input data (text vs. audio vs. video) is novel and significantly enhances the accuracy of sarcasm detection in real-world applications.
A4	Consequences to society / Environment	The project aims to improve communication for individuals with autism and those relying on assistive technologies. By enhancing sarcasm detection, it can help these individuals engage better in social and professional settings. Additionally, it could improve automated systems like customer service bots, reducing misunderstanding in AI-driven interactions. The environmental impact is minimal but could be optimized by using energy-efficient training algorithms and hardware.
A5	Familiarity	This project requires familiarity with a variety of technologies, including natural language processing tools (BERT, DistilBERT), audio processing models (Wav2Vec, AST), video processing (ViT), and multi-modal machine learning techniques. It also requires understanding of deep learning frameworks like PyTorch and TensorFlow, as well as cloud computing platforms for GPU-based training. Familiarity with ethical considerations regarding data privacy is essential, too.

Chapter 8 Conclusions

This chapter summarizes the key findings of the study, highlights the limitations encountered during the experiment, and provides suggestions for future improvements.

8.1 Summary

In this study, a multimodal approach to sarcasm detection was explored by combining three modalities: text, audio, and video. Various configurations were tested using different feature extractors, including BERT and DistilBERT for text, Wav2Vec and AST for audio, and ViT for video, with MLP and attention-based architectures for classification. The results clearly demonstrated that incorporating all three modalities—text, audio, and video—led to the best performance, with the highest accuracy of 80.02% and F1 score of 0.73 observed in the DistilBERT (text), Wav2Vec (audio), and ViT (video) combination using MLP.

The study also highlighted the relative strengths and weaknesses of using MLP versus attention mechanisms. While MLP models generally outperformed their attention counterparts, the attention mechanisms were particularly beneficial in more complex three-modal setups, allowing the model to focus on relevant features across different modalities.

Furthermore, the audio modality emerged as a key contributor to sarcasm detection. While video features also added value, audio, with its tone and pitch cues, was the most critical modality in differentiating sarcastic from non-sarcastic speech.

Overall, the findings reinforce the importance of multimodal fusion for sarcasm detection and demonstrate the effectiveness of combining MLP and attention-based architectures for this task.

8.2 Limitations

Despite the promising results, several limitations were encountered during the study. First, there was an issue with overfitting, particularly in the three-modality models, where the models performed well on training data but struggled to generalize effectively to test and validation data. This could be addressed by incorporating regularization techniques, such as dropout or data augmentation.

Additionally, the video data quality played a significant role in the model's performance. While ViT was used as the video feature extractor, the video quality (different angles of camera) could have limited the model’s ability to extract meaningful features, especially in cases where non-verbal cues were subtle. Higher quality video data could help improve the model’s performance in the future.

The attention mechanisms, while showing potential, did not consistently outperform MLP models. This suggests that attention may not be beneficial for all configurations and may require further refinement to fully leverage its potential. The complexity of attention models also contributed to longer training times and increased computational overhead, making it more challenging to optimize for efficiency.

Lastly, the imbalance in the dataset between sarcastic and non-sarcastic samples may have contributed to some of the performance issues. The model could have been biased toward the majority class (non-sarcastic), leading to suboptimal performance on the sarcastic class, which is the primary focus of the task.

8.3 Future Improvement

Several avenues for future improvement can be explored to build on the findings of this study.

1. Model Optimization: The attention mechanism showed promise, but further exploration into hybrid models that combine the strengths of both attention and MLP could yield better results. Additionally, Using techniques such as cross-modal attention or multimodal transformers could be beneficial in understanding complex relationships between text, audio, and video.

2. Data Quality: Improving the quality of the video data could significantly boost the performance of multimodal models. Higher-resolution videos with better frame rates and clear facial expressions could help the model capture subtle visual cues more effectively. Video preprocessing techniques, such as face detection, emotion recognition, or pose tracking, could be integrated to focus the model's attention on the most relevant visual features for sarcasm detection.

3. Dealing with Imbalanced Data: One key improvement would be addressing the issue of imbalanced data, where non-sarcastic samples outnumber sarcastic ones. Techniques like oversampling, undersampling, or using weighted loss functions could help mitigate this issue and ensure the model does not become biased toward the majority class. Additionally, expanding the dataset with more sarcastic samples would help balance the model's understanding of both classes.

4. Multilingual Sarcasm Detection: Although this study focused on Bengali, expanding the dataset to include multiple languages could make the model more robust and capable of detecting sarcasm in various linguistic and cultural contexts. Multilingual models could be trained to detect sarcasm across languages by leveraging multilingual BERT or DistilBERT, as well as incorporating audio features that account for language-specific tonal differences.

References

[1] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, “Towards Multimodal Sarcasm Detection (An Obviously Perfect Paper),” *arXiv preprint arXiv:1906.01815*, 2019.

[2] S. Farabi, et al., “A Survey of Multimodal Sarcasm Detection,” *in Proc. IJCAI 2024*, 2024.

[3] H. Fang, D. Liang, and W. Xiang, “Multi-modal sarcasm detection based on Multi-Channel Enhanced Fusion model,” *Neurocomputing*, vol. 578, pp. 127440, Apr. 2024.

[4] X. Gao, S. Bansal, K. Gowda, Z. Li, S. Nayak, N. Kumar, and M. Coler, “AMuSeD: An Attentive Deep Neural Network for Multimodal Sarcasm Detection Incorporating Bi-modal Data Augmentation,” *arXiv preprint arXiv:2412.10103*, 2024.

[5] T. S. Apon, R. Anan, E. A. Modhu, A. Suter, I. J. Sneha, and M. G. R. Alam, “BanglaSarc: A Dataset for Sarcasm Detection,” *arXiv preprint arXiv:2209.13461*, 2022.

[6] S. K. Lora, “A transformer-based generative adversarial learning to sarcasm detection from Bengali text based on available limited labeled data,” *Journal/Conference*, 2023. (Please replace “Journal/Conference” with actual publication info when known.)

[7] “Sarcasm detection,” *ScienceDirect Topics on Computer Science*, available at: https://www.sciencedirect.com/topics/computer-science/sarcasm-detection, accessed [Access Date].

[8] H. Xue, L. Xu, Y. Tong, R. Li, J. Lin, and D. Jiang, “Enhancing Multimodal Sarcasm Detection with Context-Aware Self-Attention Fusion and Word Weight Calculation,” *in Proc. LREC-COLING 2024*, May 2024.

[9] S. K. Singh, P. R. Patra, and R. Ghosh, “Prosodic Cues for Sarcasm Detection in Bengali,” *Journal of Speech Science*, vol. 12, no. 4, pp. 158-168, 2022.

[10] R. Ahmed, M. U. Hassan, and F. Islam, “Facial Expression Recognition for Bengali Sarcasm Detection,” *Proc. International Conference on Computer Vision*, vol. 24, no. 2, pp. 45-54, 2023.

[11] M. Anwar, Z. Li, L. Yuan, and Y. Wang, “Facial Expression Recognition for Sarcasm in Bengali: A Deep Learning Approach,” *IEEE Access*, vol. 11, pp. 4701-4714, 2023.

[12] “A systematic review of sarcasm recognition — multimodal,” *arXiv preprint arXiv:2509.04605*, 2025.

[13] T. P. Yadav, P. R. Chakrabarti, and V. Sharma, “Co-Attention Networks for Multimodal Sarcasm Detection in Bengali,” *Proc. ACM Multimedia Conference*, 2024.

[14] “Explainable Multimodal Sarcasm or Humor detection from Videos (VisTRo),” preprint, 2025. (Preprint under review.)