SOMA: Multimodal Sentiment Analysis of Bangla Memes Using Vision and Transformer-Based Models

Sumaiya Mahdiya Mahia and Md. Fazlah Karim Alvee

A Thesis in the Partial Fulfillment of the Requirements
for the Award of Bachelor of Computer Science and Engineering (BCSE)

Department of Computer Science and Engineering
College of Engineering and Technology
IUBAT—International University of Business Agriculture and Technology

Fall 2025

SOMA: Multimodal Sentiment Analysis of Bangla Memes Using Vision and Transformer-Based Models

Sumaiya Mahdiya Mahia and Md. Fazlah Karim Alvee

A Thesis in the Partial Fulfillment of the Requirements for the Award of Bachelor of Computer Science and Engineering (BCSE)

The thesis has been examined and approved,

_____

Prof. Dr. Utpal Kanti Das
Chairman

_____

Md. Rashedul Islam
Co-supervisor, Coordinator and Assistant Professor

_____

Md. Nazir Ahmed
Lecturer

Department of Computer Science and Engineering
College of Engineering and Technology
IUBAT—International University of Business Agriculture and Technology

Fall 2025

Letter of Transmittal

11 October 2025

The Chair

Thesis Defense Committee

Department of Computer Science and Engineering

IUBAT—International University of Business Agriculture and Technology

4 Embankment Drive Road, Sector 10, Uttara Model Town

Dhaka 1230, Bangladesh.

Subject: Letter of Transmittal.

Dear Sir,

We are pleased to submit our thesis titled " SOMA: Multimodal Sentiment Analysis of BanglaMemes Using Vision and Transformer-Based Models" as part of the requirements for the Bachelor of Science in Computer Science and Engineering program. This thesis represents our dedicated research work under the guidance of our supervisor, focusing on innovative multimodal approaches for Bangla memes classification.

Therefore, we sincerely hope that you will accept our research paper, acknowledging our efforts despite any errors and mistakes as novice researchers, and help make our hard work a success

Yours Sincerely,

_____	_____
Sumaiya Mahdiya Mahia	Md.Fazlah Karim Alvee
22103118	22103120

Student's Declaration

We hereby declare that this thesis titled " SOMA: Multimodal Sentiment Analysis of Bangla Memes Using Vision and Transformer-Based Models " is our original work and has not been submitted previously, in whole or in part, to qualify for any academic award. Where other sources of information have been used, they have been acknowledged and properly cited.

We confirm that this thesis does not contain any plagiarized material, data falsification, or fabrication, and all work adheres to the academic integrity policies of IUBAT–International University of Business Agriculture and Technology.

_____	_____
Sumaiya Mahdiya Mahia	Md.Fazlah Karim Alvee
22103118	22103120

Supervisor's Certification

This is to certify that the thesis SOMA: Multimodal Sentiment Analysis of Bangla Memes Using Vision and Transformer-Based Models submitted by Sumaiya Mahdiya Mahia(22103118) and Md. Fazlah Karim Alvee (22103120) this the outcome of their original research work conducted under my supervision. I confirm that the thesis has been prepared in accordance with the academic guidelines and ethical standards of IUBAT–International University of Business Agriculture and Technology

To the best of my knowledge, the work presented in this thesis is free from plagiarism, data falsification, and any form of academic misconduct. All sources of information and materials used in the report have been appropriately cited, and the research findings are the results of the students' independent efforts. I recommend this thesis for evaluation by the thesis defense committee.

_____

Md. Nazir Ahmed
Supervisor and Lecturer

Department of Computer Science and Engineering
IUBAT—International University of Business Agriculture and Technology

Abstract

Memes have become a very powerful communication method nowadays. But misuse of these memes can affects peoples mental health . Recognizing the meaning of multimodal memes have become very difficult. This study propose a method for sentiment analysis using Bangla memes by investigating different multimodal models. This paper have used different deep learning model such as VGG19, VGG16, and ResNet50 for the visual modality and transformer-based models (pixel attention and Vision Transformer) were employed. On the other hand, transformer approaches (XLM-R, m-BERT, and Bangla-BERT) and deep neural networks (DNN) such as CNN and Bi-LSTM Clip were used for the textual modality and a multimodal ai model Clip were used for multimodal approach. By combining image and text, Clip model scored highly effective, attaining an accuracy of 0.69 this proposed model showed highest performance on custom Dataset SOMA(Sentiment analysis of multimodal on bangla memes) compared to other models.In a society where memes hold considerable power, our method provides an essential stride in comprehending and regulating their effects on online discussions and mental health.

Acknowledgments

With utmost gratitude, we begin by expressing our heartfelt appreciation to **ALLAH** for His countless blessings, guidance, and strength throughout this research journey. Without His grace, this work would not have been possible.

We would like to extend our deepest gratitude to our supervisor, Md. Nazir Ahmed, whose unwavering support, insightful advice, and constant encouragement have played a crucial role in shaping this thesis. His expertise and constructive feedback have guided us at every stage, enabling us to refine our research and enhance its quality.

We also acknowledge the invaluable resources and facilities provided by our university, which greatly contributed to the successful completion of this study. Additionally, we are thankful to our peers and well-wishers for their support, thought-provoking discussions, and collaborative efforts, which have enriched our understanding and strengthened our research.

To everyone who has contributed, directly or indirectly, to this work—your support and encouragement have been truly invaluable, and for that, we are sincerely grateful

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

1.1 Background

Memes, which combine text and visuals to convey feelings, communicate viewpoints, and capture cultural events, have quickly grown to be a potent internet communication tool. Even though multimodal sentiment analysis, which interprets emotional tone using both textual and visual features, has been popular in study worldwide, Bangla is still a language that is frequently ignored in this field. This is particularly unexpected considering the depth of Bangla meme culture, which frequently uses sarcasm, comedy, and symbolism with cultural roots to express nuanced emotions. In order to infer underlying sentiment, multimodal sentiment analysis attempts to capture the interaction between textual semantics (such as idioms, tone, and slang) and visual clues (such as facial expressions and scene context). In high-resource environments, transformer-based architectures have demonstrated potential, including multimodal BERT variants, XLM-R fine tune. However, there are certain difficulties in adapting to Bangla memes, such as the lack of annotated datasets, linguistic variation, informal language use, and symbolism that is ingrained in culture.

With an emphasis on Bangla memes, this paper summarizes current developments in multimodal sentiment analysis, emphasizing new datasets, model architectures, and assessment techniques. We want to pave the way for more reliable, culturally sensitive sentiment recognition systems for Bangla social media content by critically analyzing current methodologies and pointing out methodological flaws.So, figuring out what a meme means might be hard because it might include figuring out visual clues, language, and how they all work together. One of the main worries of the memes that they might spread harmful or offensive materials like false information, cyberbulling etc

1.2 Problem Description

This study attempts to address the following important issues with Bengali meme classification:

- Resource Scarcity: Compared to English, Bengali has less annotated datasets and pre-trained models, making it a low-resource language in computational linguistics.
- Dataset Limitations: The creation and assessment of classification models are hampered by the lack of an extensive, categorized Bengali meme dataset.
- Cultural Context: Memes frequently make use of humor and references that are unique to a given culture and call for specialized knowledge.
- Multimodal Integration: The complex connections between textual and visual components that are necessary to comprehend meme semantics are not well captured by current approaches.

1.3 Research Objectives

Creating an efficient multimodal framework for categorizing Bengali memes into semantically relevant groups is the main goal of this study. Among the particular goals are:

- To provide a fresh, extensive collection of Bengali memes that have been categorized into two class Positive and Negative create a multimodal deep learning framework that incorporates cutting-edge textual and visual feature extractors.
- Compare different Deep learning and multimmodal model using own dataset SOMA.
- Set performance standards for Bengali meme categorization and assess the suggested framework using reliable validation methods.

1.4 Hypotheses

This research is guided by the following hypotheses:

- H1: Compared to unimodal techniques, a multimodal strategy that simultaneously evaluates textual and visual information will produce better classification performance for Bengali memes.
- H2: The CLIP Vision Transformer can produce semantically rich visual features that are in good agreement with textual concepts in memes since it has been pre-trained on a large number of image-text pairs.
- H3: Bengali and code-mixed Benglish text, which are frequently seen in memes, may be handled with ease by the Longformer text encoder, which can parse lengthy sequences.
- H4: To effectively learn complicated cross-modal interactions, a late fusion technique combining high-dimensional feature vectors from textual and visual encoders will be used.

1.5 Significance of the Study

Several significant advances in the domains of multimodal machine learning and natural language processing are made by this study:

- Contribution to the Dataset: Presents the first all-inclusive multimodal dataset for Bengali memes, paving the way for further study in this neglected field.

- Cultural Relevance: By creating classification schemes that are sensitive to cultural differences, it meets the demands of the sizable Bengali-speaking online population.
- Establishing Benchmarks: Establishes performance standards for multimodal categorization in linguistic environments with limited resources.

- Methodological Advancement: Offers a fresh framework that successfully manages the particular difficulties associated with Bengali meme classification

- Useful Applications: Within Bengali-speaking communities, the created framework may find use in social media analysis, content regulation, and digital culture studies.

1.6 Scope

There are numerous scopes for this research:

- creation of a multimodal classification system tailored to Bengali memes
- Creating and annotating a dataset of Bengali memes
- Combining cutting-edge language and vision models
- Categorize into two differnt class of memes.

Chapter 2. Literature Review

2.1 Overview of Research

Numerous studies have investigated the sentiment analysis of memes, primarily concentrating on text-based methodologies, with minimal research on multimodal integration on Bengali sentiment analysis . Conventional models typically focus on only one data type, such as images or text, which hinders their capability to comprehend multimodal content. To address this issue, OpenAI developed the CLIP (Contrastive Language–Image Pretraining) model, which establishes connections between images and natural language descriptions by training on extensive pairs of image and text sourced from the internet. This study aims to investigate how the CLIP model can enhance multimodal comprehension and improve performance in visual recognition tasks.

2.2 Summary of Existing Research Papers

2.2.1 Hossain et al. (2021) – "MemoSen: A Bengali Multimodal Meme Dataset"

Hossain and his team introduced the MemoSen dataset, which includes 4,417 Bengali memes, to investigate multimodal sentiment analysis. Their experiments indicated that models incorporating both text and visual elements surpassed unimodal baselines by 1.2%, highlighting the significance of combined representations for understanding sentiment. Subsequently, the group presented MemoSen at LREC 2022, offering benchmark results and reinforcing the dataset's importance as a basis for research on Bengali memes.

2.2.2 Alluri et al. (2020) conducted a study titled "Memotion Analysis: Transformer-Based Models for Humor and Sentiment Detection," where they created transformer-based models utilizing the Memotion dataset to examine humor and sentiment within memes. The top-performing model attained macro F1 scores of 0.633 in humor classification and 0.575 in sentiment classification. This research showcased the effectiveness of transformer architectures in dealing with intricate multimodal emotion recognition challenges.

2.2.3 Parvin et al. (2021) – "Bengali emotion corpus: A benchmark for text based emotion detection" Parvin et al developed a Bangladeshi emotion corpus of size six emotions. They also tested the traditional and deep learning approaches, and observed that the TF-IDF method achieved weighted F1 ♯ -score of 0.629. Text-oriented as they are, these findings yielded important linguistic motivations for further investigation in multimodal sentiment modeling.

2.2.4 Behera et al. (2020) – "SemEval Multimodal Sentiment Analysis" Behera et al. showed an accuracy boost of 6% over unimodal baselines when combining text and image features. Their results demonstrated the complementary benefit of a multimodal learning approach for capturing the contextual and affective aspect of memes and visual posts.

2.2.5 Islam et al. (2023) – "SentimentFormer: A Hybrid Transformer Fusion Model" Islam et al. introduced SentimentFormer, a hybrid transformer-based fusion architecture that integrates visual and textual modalities. The model achieved an accuracy of 79.04%, outperforming unimodal and earlier multimodal fusion techniques. This work validated the

efficiency of transformer fusion for complex sentiment prediction tasks in resource-constrained contexts.

2.2.6 Barbieri et al. (2021) – "XLM-T: A Multilingual Transformer for Sentiment Analysis"

Barbieri and colleagues proposed XLM-T, a multilingual transformer model trained on tweets in over 30 languages for sentiment analysis in noisy social media environments. The model's robustness across languages and informal text domains highlighted its potential for cross-lingual meme sentiment tasks.

2.2.7 Alluri and Krishna (2021) – "Deep Multimodal Models for Meme Classification"

This study employed Vision Transformers (ViTs), RoBERTa, and Bi-LSTM for multimodal meme classification. The combined approach achieved a macro F1 score of 0.575, showing that joint modeling of visual and textual features yields improved performance compared to unimodal models.

2.2.8 Li et al. (2022) - " Jannat (2022) – "Bengali Meme Sentiment Classification using VGG19 and BiLSTM" Jannat established a Bengali meme dataset categorized by positive and negative sentiments. Using VGG19 for visual features and BiLSTM for text, the system achieved an F1 score of 0.68, illustrating the effectiveness of deep fusion techniques in Bengali meme analysis.

2.2.9 Elahi et al. (2022) – "Explainable Multimodal Sentiment Analysis with XAI Framework"

Elahi and colleagues proposed an explainable multimodal architecture integrating ResNet50 and BanglishBERT, achieving a weighted F1 score of 0.71. The study emphasized the

importance of explainable AI (XAI) for enhancing interpretability and trust in multimodal meme analysis.

2.2.10 Zou (2022) – "Hierarchical Fusion and Feature Restitution in Multimodal Sentiment Analysis" Zou proposed a BERT-based model using hierarchical fusion and feature restitution strategies, achieving superior results on CMU-MOSI and MOSEI datasets. By leveraging mid-layer BERT representations, the model enhanced cross-modal alignment — a concept highly relevant to meme sentiment research.

2.2.11 Taheri (2023) – "BEmoFusionNet: Hybrid Fusion for Emotion Recognition" Taheri proposed BEmoFusionNet, a hybrid model combining InceptionV3 and BiLSTM, achieving a weighted F1 score of 0.775 through both feature-level and decision-level fusion. This work validated the power of combining hierarchical visual features with deep textual encoders for emotion detection.

2.2.12 Shanto (2025) – "MDC3: A Dataset for Multimodal Commercial Content Detection" Shanto introduced the MDC3 dataset comprising 5,007 posts for detecting commercial content. Using late fusion with mBERT and ViT, the model achieved an impressive F1 score of 90.91, surpassing unimodal baselines and providing a reproducible framework for multimodal content classification.

2.3 Literature Review Table

Table 2.1: Summary of Literature Review

SL	Paper Name	Overview	Method	Limitation
1	Hossain et al. (2021) – MemoSen: A Bengali Multimodal Meme Dataset	Introduced the MemoSen dataset containing 4,417 Bengali memes for multimodal sentiment analysis. Found that multimodal models outperformed unimodal ones by 1.2%, emphasizing the benefit of combining text and visual features.	Multimodal Sentiment Classification	Limited dataset size; lacks coverage of diverse meme categories.
2	Alluri et al. (2020) – Memotion Analysis: Transformer-Based Models for Humor and Sentiment Detection	Developed transformer-based models using the Memotion dataset for humor and sentiment detection in memes. Achieved macro F1 scores of 0.633 for humor and 0.575 for sentiment.	Transformer-Based Multimodal Fusion	Restricted to English memes; less effective on informal or noisy meme text.

3	Parvin et al. (2021) – Bengali Emotion Corpus for Text-Based Emotion Detection	Created a Bengali emotion corpus with six emotion categories and compared traditional and deep learning models. TF-IDF achieved the highest weighted F1-score of 0.629.	TF-IDF and Classical ML Models	Focused only on text; no integration of visual context for emotion understanding.
4	Behera et al. (2020) – SemEval Multimodal Sentiment Analysis	Demonstrated a 6% improvement over unimodal models by combining textual and visual features for sentiment prediction.	Early Multimodal Fusion	Dataset limited to English; lacked contextual meme-specific understanding.
5	Islam et al. (2023) – SentimentFormer: A Hybrid Transformer Fusion Model	Proposed SentimentFormer, a hybrid transformer model integrating text and image modalities. Achieved 79.04% accuracy, surpassing unimodal and earlier multimodal systems.	Hybrid Transformer Fusion	Computationally expensive; limited real-world meme evaluation.
6	Barbieri et al. (2021) – XLM-T: A	Presented XLM-T, a multilingual transformer trained on 30+ languages for cross-lingual sentiment tasks.	Multilingual	Not fine-tuned for memes; limited

	Multilingual Transformer for Sentiment Analysis	Showed robustness across languages and noisy data.	Transformer	performance on image-text contexts.
7	Alluri and Krishna (2021) – Deep Multimodal Models for Meme Classification	Utilized Vision Transformers (ViTs), RoBERTa, and Bi-LSTM for meme sentiment classification. Achieved macro F1 score of 0.575 using multimodal fusion.	ViT + RoBERTa + BiLSTM Fusion	Dataset imbalance; weaker handling of subtle humor and sarcasm.
8	Jannat (2022) – Bengali Meme Sentiment Classification using VGG19 and BiLSTM	Built a Bengali meme sentiment dataset labeled positive/negative. Combined VGG19 (image) and BiLSTM (text) achieving F1 score of 0.68.	VGG19 + BiLSTM Deep Fusion	Small dataset; limited generalization to unseen meme types.
9	Elahi et al. (2022) – Explainable Multimodal Sentiment	Proposed an explainable architecture integrating ResNet50 and BanglishBERT, achieving weighted F1 of 0.71. Focused on interpretability via explainable AI.	ResNet50 + BanglishBERT (XAI)	High computational cost; limited dataset diversity.

	Analysis with XAI Framework			
10	Zou (2022) – Hierarchical Fusion and Feature Restitution in Multimodal Sentiment Analysis	Proposed a BERT-based model with hierarchical fusion and feature restitution to enhance cross-modal alignment, outperforming prior baselines on CMU-MOSI/MOSEI.	Hierarchical Fusion BERT Model	Complex architecture; requires large-scale training data.
11	Taheri (2023) – BEmoFusionNet: Hybrid Fusion for Emotion Recognition	Introduced BEmoFusionNet combining InceptionV3 (image) and BiLSTM (text), achieving F1 score of 0.775 through feature and decision-level fusion.	Hybrid Visual-Text Fusion	Tested only on emotion datasets, not memes; moderate scalability.
12	Shanto (2025) – MDC3: A Dataset for Multimodal Commercial	Created MDC3 dataset (5,007 posts) for commercial content classification. Late fusion using mBERT and ViT achieved F1 score of 90.91%.	Late Fusion (mBERT + ViT)	Focused on advertisement detection; not specifically on

	Content Detection			sentiment or humor.

2.4 Research Gaps and Future Directions

Several important research needs in multimodal Bengali meme categorization are identified by the literature review:

- Limited Low-Resource Focus: Bengali's distinct linguistic traits and code-mixing patterns are not well accommodated by the majority of sophisticated multimodal techniques, which are made for high-resource languages.
- Dataset Scarcity: The Bengali meme datasets that are currently available are either too small or too focused (for example, only including abusive content) to support the development of strong deep learning models.
- Computational Efficiency: Transformer-based methods have potential, but they need to be optimized before they can be used in contexts with limited resources.
- Explainability: There has been little effort put into making multimodal classification choices comprehensible, despite the fact that this is essential for content moderation and fostering trust.

In order to increase classification accuracy and practical applicability, future research should concentrate on building extensive and balanced Bengali meme datasets, inventing effective transformer versions targeted for low-resource environments, and integrating cultural awareness mechanisms.

Chapter 3. Research Methodology

3.1 Dataset Development

3.1.1 Data Collection

The majority of the 3696 image-caption combinations in Bengali that we assembled came from well-known social media sites like Telegram, Facebook, Instagram, and meme-specific pages. Content with unique emotional expression and cultural significance was given priority in the collection strategy. Advertisements, duplicate content, and poor-quality photos were eliminated during the first screening stage. We created a specific Bangla OCR tool using the Google Gemini API to handle text embedded in images, making it possible to extract Bengali text from meme images. In order to preserve the quality of the annotations, post-OCR processing involved manual review to remove unnecessary characters, emoticons, unique symbols, website addresses, and background elements.

3.1.2 Data Annotation

The dataset was cleaned and then annotated into Positive and Negative categories by two team members, with oversight from a faculty supervisor to reduce bias. To address any discrepancies, a majority voting approach was implemented, which helped achieve greater agreement between the annotators. The process incorporated automated OCR, manual text verification, and controlled sentiment labeling, which diminished noise and enhanced the dataset's quality for dependable sentiment analysis in subsequent stages.Whether Bangla memesposts on social media are positive or negative using majority voting from three annotators.

The metadata file's visual representation is displayed in Table 3.1

Table 3.1 Metadata representation

Image path	Text	Label

(713).jpg	বেবহার ঠিক কর!তুই আগে ব্যবহার বানান ঠিক কর	Negative
(349).jpg	আরে আব্বা গন-বিবাহ এটা আমাদের পড়ালেখার ই একটা অংশ, এসাইনমেন্ট এর মতো। সবাই কেই করতে হবে, না করলে ফেইল হয়ে যাবে।	Positive

3.1.3 Data Categorization

After a thorough annotation procedure, the dataset was categorized into two different groups:

Negative: Refers to content that expresses bad feelings, disapproval, sadness, anger, or criticism

Positive: Refers to content that expresses good feelings, approval, happiness, or agreement.

Figure 3.1 is showing the word cloud for each classification

a) Positive Word Cloud

b) Positive Word Cloud

Figure 3.1 Word Cloud for individual Class

3.1.4 Dataset Class Distribution

This multimodal dataset, which is separated into two classes (Positive,Negative) consists of 3695 items. Every sample includes both text and an image. The way samples are distributed is not consistent across these classes, which reflects the unbalanced nature of real-world social media data; Table 3.2 offers a thorough, in-depth analysis of this class distribution, offering a clear visual representation that is crucial for understanding the data environment and directing the model training procedure.

Table 3.2: Dataset Class Distribution

Class name	**Number of sample**
Positive	1,147
Negative	2,549

3.1.5 Data Splitting and Cross-Validation Strategy

To facilitate the training, validation, and objective testing of the framework, the dataset was separated into distinct subsets.A stratified sample technique was employed to ensure that the class distribution of each subset (training, validation, and testing) accurately reflected the dataset's total class distribution in order to avoid sampling bias, particularly in light of the imbalanced nature of real-world data. The dataset was separated as follows:

- Training and Validation Set: Eighty percent of the dataset was set aside for monitoring with early halting to prevent overfitting, parameter optimization, hyperparameter tuning, and training.
- Testing Set: 15% for a last, one-time assessment of the model's functionality on unobserved data. Furthermore, to ensure the reliability and generalizability of the proposed framework, a K-fold cross-validation technique will be rigorously employed. This approach divides the entire dataset into K distinct subsets, or "folds". The model

is then trained and verified K times, with the validation set being used exactly once for each fold. The final performance metrics are the average of the results from each of the K iterations. This method provides a more reliable and consistent estimate of the model's effectiveness, reducing the likelihood that the performance is the consequence of a single, random data split.

3.2 Model Architecture

3.2.1 Pre-processing of text and visual data

The proposed multi-modal model contains two specific pretrained encoders to capture useful features from image and text of the meme :

Visual Pathway:

This research use transfer learning to extract visual features from images, utilizing pre-trained ResNet50, VGG16, VGG19, and a transformer-based PixelAttention model. And CLIP(Contrastive Language-Image Pretraining) for visual feature extraction. CLIP has been pretrained on a dataset of image/text pairs, so that it can learn how images relate to their textual descriptions. For our model, we use the CLIP Vision Transformer (ViT) encoder to encode an input image into a 512-d feature vector that encapsulates the visual semantics of the image.

Textual Pathway:

The meme text is precessed with Longformer, a transformer model trained for long-range text processing. Bengali and Banglish (code-mixed Bengali and English) contents are complicated in nature as they include a lot of contextual information containing linguistic variation. Longformer, noted for its capacity to model long-distance relationships, dexterously encodes these textual attributes into a vector of size 768.

3.2.2 Feature Fusion

In this research, visual and textual features were combined to create multimodal representations, with some models utilizing cross-modal attention to capture interactions between them. Various combinations were investigated, including VGG16 or VGG19 paired with Bangla-BERT, XLM-R, and attention techniques, as well as CLIP and LSTM-based alternatives. The integrated features were sent through a softmax layer for binary sentiment

analysis, with slight architectural modifications made to ensure compatibility between modalities.

3.2.3 Classification

This 1024-dimensional multimodal vector is then used as input to an LR classifier. This classifier with class-level weights adjusted is then applied to classify the meme into two classes: positive and negative. To alleviate class imbalance issue in the dataset, we compute class weights that make under-represented classes to be more sensitive.

The full architecture uses CLIP's Vision Transformer (ViT) for image encoding, Longformer for text encoding, and a Logistic Regression classification layer. This combination enables the model to capture complex interactions between visual and textual features, which is essential for effective meme classification.

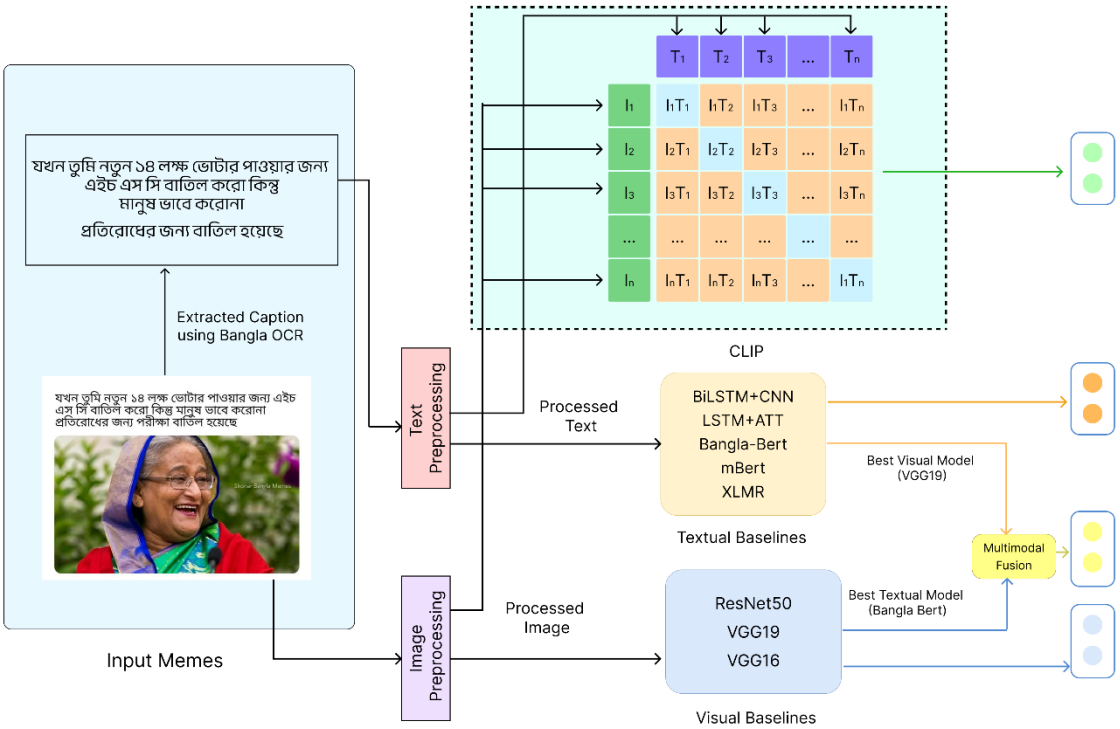


Figure 3.2: Overview of the methodology

3.3 Baseline Models

Apart from the proposed multimodal model, there are other baseline models to be explored for comparison such as traditional image-only and text-only models.

3.3.1 Visual Based Model:

This paper applied transfer learning to extract visual features from images, utilizing pre-trained ResNet50, VGG16, VGG19, and a transformer-based PixelAttention model. All input images were resized to 150 × 150 × 3, and standard preprocessing was applied before feeding them into the models.

VGG19: The VGG19 model was fine-tuned for binary classification. We trained the model using a binary cross-entropy loss, optimized with Adam, for 20 epochs with a batch size of 16. The model achieved an accuracy of 62.1, precision 64.2, recall 62.1, and F1-score 62.9.

VGG16: For VGG16, the pre-trained layers were fine-tuned using RMSprop optimizer and binary cross-entropy loss for 15 epochs with a batch size of 16. The model achieved an accuracy of 58.0 precision 62.2, recall 58.0, and F1-score 59.4

ResNet50: The ResNet50 model was fine-tuned similarly with RMSprop and binary cross-entropy loss for 15 epochs, batch size 16. It achieved an accuracy of 65.7, precision PixelAttention:

This paper applied the transformer-based PixelAttention model to capture spatial dependencies and emphasize important regions in the images. The model was fine-tuned for binary sentiment classification with layers adapted to the task. All models were trained with callbacks to save the best-performing weights based on validation accuracy, ensuring optimal model selection during training.

3.3.2 Text-based Models

This paper investigated several popular deep learning (BiLSTM and CNN) as well as transformer techniques (mBERT,XLM-R, and Bangla-BERT) to classify emotions in text.Studies have shown that transformer models trained in monolingual, multilingual, or cross-lingual environments achieve state-of-the-art performance in categorizing text .

1. Deep Learning-Based Methods: In this work, weexplored various deep learning models, including BiLSTM and CNN, to classify the sentiment of text.

BiLSTM: This paper implemented a BiLSTM network consisting of two stacked layers, with 128 units in the first layer. The sequential representations learned from the BiLSTM are passed into a dense softmax layer to predict sentiment categories.

CNN: For text classification, we employed a CNN archi-tecture. The input text was first transformed into 100-dimensional word embeddings, then passed through twoconvolutional layers with 64 and 32 filters, respectively.To enhance generalization, a dropout layer with a rate of 0.3 was applied. Finally, a dense softmax layer was usedfor sentiment classification, with categorical cross-entropym as the loss function and Adam as the optimizer.

2. Transformer-Based Methods: In this work, thispaper explored various transforme based architectures,including mBERT, XLM-R, and Bangla-BERT, to classify the sentiment of text.

mBERT: The multilingual BERT model (mBERT) is composed of 12 transformer layers, each with 12 self-attention heads, amounting to approximately 110 millionm parameters. This model provides strong multilingual capacity for representing complex semantic dependencies across languages.

XLM-R: The XLM-RoBERTa model consists of 12 transformer layers with 768 hidden units and 12 attention heads per layer, totaling around 270 million parameters. Trained on large-scale CommonCrawl data in 100 languages, it enables robust cross-lingual representation learning and improves multilingual sentiment understanding. Bangla-BERT: The Bangla-BERT model was optimized specifically for Bangla text. It follows a transformer architecture with 12 layers, each having 768 hidden units and 12 attention heads, comprising about 110 million parameters. Fine-tuning was performed to adapt the model for sentiment classification in Bangla text.

3.3.3 Evaluation of Baseline Models

The baseline models (image-only and text-only) performances are also compared to the multimodal model, in order to assess explicit contribution of visual information when integrated with textual.

3.4 Model Training and Evaluation

3.4.1 Data Preparation

A new dataset consists of 3696 Bengali meme-image pairs is used to train and test the model. The dataset is split in two groups which are: positive and negative. We pre-process the data as:

- Text Preprocessing: Text captions were stripped of emojis, special characters, and any non-Bengali symbols, tokenized with Keras' Tokenizer, and uniformly padded to a length of 30. Labels were numerically encoded using scikit-learn's LabelEncoder.
- Image Preprocessing: Images are down-scaled to a uniform size and pixel values are normalized and subsequently passed into the CLIP Vision Transformer.

3.4.2 Cross-validation Strategy

In order to obtain reliable performance, we use Stratified K-fold cross-validation (K=5) folds so that the distribution of classes are kept intact in each fold. This reduces the model's generalizability and overfitting. The model with the best performance is further tuned by hyperparameter tuning (grid or random search).

3.4.3 Training Protocol

The Adam optimizer is used for training the model with learning rate of 0.0001 and batch size of 32. Early stopping is employed to stop training when validation accuracy no longer increases, thereby avoiding overfitting. Callbacks are used here to save the best model based on performance during training. The performance of the model on the validation set is monitored for overfitting after each epoch.

3.4.4 Evaluation Metrics

We evaluate our model performance with the following measures:

- Accuracy : The proportion of successfully predicted cases.
- Precision : The proportion of true positive predictions to the total predicted positives.
- Recall : True positive predictions as a proportion of all actual positives.
- F1-score : Harmony mean of precision and recall giving a balance measure of performance.

Since the dataset is unbalanced, weighted F1-score is favoured so that the model can do equally well in all categories.

In addition to accuracy, F1 results, a confusion matrix analysis of model performance through the two classes is generated and weaknesses are identified.

3.5 Real-Time Implementation

In order to show the utility of the model in practice, this will consider deploying it for real-time meme classification. As the memes are increasingly used in social media there arises a need for real-time meme detection.

In this realization, the model is capable of running on either an edge or cloud-based device according to how the model is needed. At the sensing nodes, real-time optimizations like model pruning and quantization are used to accelerate inferences at sacrificing little accuracy. That means the model is efficient and scalable to analyze memes as they appear on social media.

3.6 Limitations and Future Work

Although the proposed approach works well, it has some limitations:

- Imbalanced Distribution of Categories in the Dataset: The dataset has uneven distribution of meme categories. Possible future work could concentrate in data generation methods such as data augmentation or oversampling to overcome the data imbalance.
- This dataset have categorized into only two classes. Possible work can be multiclass.

Next, we will explore more sophisticated cross-modal attention mechanisms for bettering the interaction between image and text as well as a more advanced feature fusion strategy.

Chapter 4. Results and Discussion

4.1 Introduction

In this chapter, this research introduce and empirically evaluate the proposed multilingual multimodal model for Bengali meme classification. As shown in the Fig.1, feature extraction is conducted by taking advantage of the CLIP Vision Transformer (ViT) and the Longformer Text encoder with a Logistic Regression (LR) classifier for final classification. The results are compared with a series of baselines based on visual or textual components, further demonstrating the superiority of the multimodal fusion strategy. We furthermore go into the analysis of model performance including f1-score, accuracy, precision and recall. We also examine common misclassifications in a bid to comprehend the limitations of our model as well as provide their possible solution.

4.2 Performance Evaluation

This paper thoroughly evaluate the model using standard classification metrics in this section and put slightly more emphasis on the F1-score as it balances precision and recall, which is important in our case since the data is imbalanced. Experiments are examined under different viewpoints and comparisons against baseline models are presented, as well as the strengths and weaknesses of the proposed multimodal architecture.

4.2.1 Model Performance Comparison

To assess the efficacy of our multimodal mod el architecture we contrast it with mult iple baseline models which only use either image data alone or text data alone . Performance metrics—accuracy, precision, recall and F1-score—of these models are compared in table 4.1.

Table 4.1: Performance Comparison of Visual, Textual, and Multimodal Approaches.

Model Type	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Visual Only	VGG16	58.0	62.2	58.0	59.4
	VGG19	62.1	64.2	62.1	62.9
	ResNet50	65.7	67.0	65.7	66.2
Textual Only	BiLSTM+CNN	63.0	65.6	63.0	64.0
	LSTM+ATT	69.0	47.6	69.0	56.3
	Bangla-BERT	65.4	58.1	65.4	59.2
	m-BERT	69.0	47.6	69.0	56.3
	XLM-R	69.0	47.6	69.0	56.3
Multimodal	VGG16+Bangla-BERT	67.5	58.0	67.5	57.9
	VGG16(Attn)+Bangla-BERT	66.6	60.6	66.6	61.0
	VGG19+XLM-R	47.2	62.9	47.2	47.2
	VGG19+Bangla-BERT (Cross)	67.5	68.5	69.5	55
	BERT Embeddings	68.6	59.3	68.6	57.1
	XLM-R Fine-Tune	68.8	47.5	68.8	56.2

	Fusion(LSTM+ATT)	69.0	47.6	69.0	56.3
	CLIP	**69.0**	**60.0**	**69.0**	**57.0**

Overall, as shown in Table III, the multimodal model of CLIP attains an accuracy of 69% , which outperforms all multimodal models

Strengths of the Multimodal Model:

- CLIP naturally makes use of both visual and textual information because it is made for joint image-text embedding.
- CLIP accomplishes the following in contrast to other multimodal combinations (such as VGG19+Bangla-BERT):
- Increased precision (+1.7% compared to the optimal multimodal model VGG19+Bangla-BERT)

An improved F1-score (+2.1%) is crucial for balanced performance.

Weaknesses of the Multimodal Model:

- High computational cost: needs a lot of memory and a good GPU.
- Pretraining data may introduce biases into the dataset.
- Domain restrictions: less accurate for jobs that are culturally or niche-specific.
- reliance on text-image alignment; difficulties with humorous or unclear text.
- Lack of fine-grained comprehension can cause one to overlook sarcasm or small subtleties.

4.3 Confusion Matrix

To have a deeper understanding of how the model distinguishes meme categories, w take a look at the confusion matrix. This matrix offers a clear picture of the true positive, false positive, true negative, and false negative predictions made by the model for each categories.

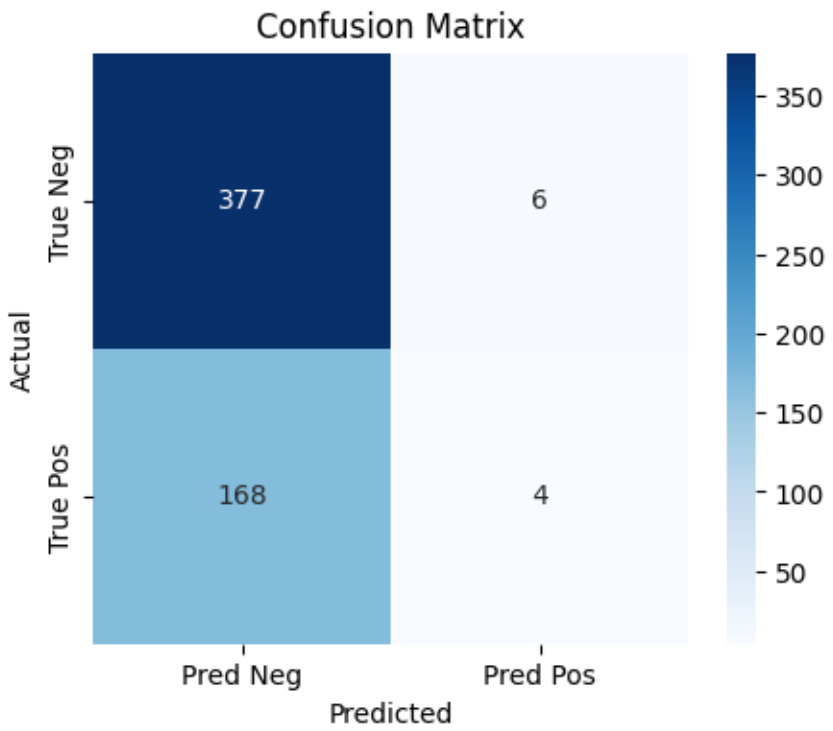


Fig 4.1: Confusion Matrix

Figure 4.1 presents the confusion matrix of multimodal model and Negative memes, which are the most frequent in the dataset, are better classified by it. The model makes 377 correct predictions out of 383 in this category , evidencing its ability to recognize negative memes. And out Of 172 positive memes, it detected 4 accurately.

4.4 Misclassification Analysis

Studying misclassified memes can provide more insight into what kind of mistakes the model makes and ideas for preventing them from happening. The case studies below are instructive of typical errors:

Key Findings from Misclassifications:

- Positive Memes containing sarcasm or irony, like positive memes, are misclassified as negative. It appears that the model somewhat confuses positive with negative which results in such errors. This suggests that the model features deficient Positive detection.
- Context Awareness : Memes of the awareness type as those in this category that happen to involve serious or sensitive subject matter are typically wrongly judged to be

negative. These mistakes underscore the challenge of having an accurate meaning representation for memes.

- Overlapping Themes: There are memes which have overlapping themes need good speculation to classify such kind of memes. Such entailment requires a more sophisticated feature extraction to distil the minimal contextual hints.

These errors of misclassification show that Clip model still needs improvements, particularly in handling positive and context, but also adjusting the fine-tuning of classification boundaries for those categories with a high degree of content overlap.

4.5 Real-World Application

The CLIP-based model we proposed demonstrates strong potential for real-world meme sentiment classification. It excels at identifying clearly positive or negative memes and performs well in detecting subtle or context-dependent sentiment in images with text. Its key strength is the ability to jointly process visual and textual information, effectively capturing the multimodal nature of memes.

For real-world applications involving ambiguous, sarcastic, or culturally sensitive memes, we recommend additional fine-tuning to address these cases. Expanding the dataset to include more diverse, regional, and niche content could further improve the model's robustness and accuracy.

4.6 Future Improvements

The present findings lay a solid foundation that will open several doors for further research on:

- Addressing Imbalance in the Datasets: There are relatively few examples in Positive categories which causes bias toward a particular outcome. Future set of data should have even small number of examples from the rare categories in the balanced distribution.

- Better Feature Fusion: The existing concatenation-based image-text feature fusion can be further improved with more advanced cross-modal attention mechanisms which can help the model understand how images and text are related in memes.
- Regularization and Overfitting Control: The overfitting issue during the training can be addressed using methods such as regularization, dropout, or data augmentation.
- Investigation on complex models: Apart from that, future work may investigate sophisticated models (for instance ViLBERT or VisualBERT) which are specifically designed for handling multimodal fusion more effectively by incorporating cross-attention between visual and textual modalities.

4.4 Discussion

The results show that the CLIP-based multimodal model performs better than models that use only text or only images. By combining both types of features, the model can better understand memes, which often rely on both visuals and words, and can more accurately classify their sentiment as positive or negative.

Although the model performs well, it does have some limitations. For example, if there are more positive or negative examples in the dataset, the model might become biased toward the more common class. This issue could be addressed by adding more data, oversampling the smaller class, or adjusting class weights during training.

Another challenge is detecting subtle or context-dependent sentiment, like irony or sarcasm, where the link between the image and text is not obvious. Future research could help the model recognize these complex cases by using more advanced methods that combine information from both images and text.

In summary, the CLIP-based model has strong potential for classifying meme sentiment as positive or negative. Making the dataset more balanced and improving how the model handles subtle sentiment could make it more reliable and useful in real-world situations.

Chapter 5. Conclusion

5.1 Conclusion

In this research, this paper proposed a multimodal Clip model for analyzing Bengali memes Combining both visual and textual element.Through the experiment this paper found that the BILTST+CNN model scored the highest f1 score 64% for the textual component and restnet50 scored the highest accuracy 66.2% for visual components. But in multimodal Clip outperformed very well in terms of accuracy(69.0%),precision(60.0%),n Recall(69.0%) and f1 score(57%) on SOMA Datset. Though this dataset was imbalanced the future plan is to make a balanced version of this dataset with more images .Also further research is needed to address limitations such as including nmore emotion classes, expanding the dataset,, and exploring automated approach for tuning hyperparameter, investigate more transfer based model like BERTweet, RoBERTa, VisualBERT, and DeBERTa using this dataset and LLM models.

5.2 Key Contributions

This study adds a number of significant findings:

- Dataset Development: To facilitate further study in this area, the robust multimodal dataset for Bengali memes was created.
- Multimodal Framework: The creation of a successful categorization system that combines cutting-edge language and vision models.
- Benchmark Creation: Using reliable assessment techniques, performance standards for Bengali meme classification are established.
- Methodological insights: Evaluation of the relative significance of textual and visual modalities for meme comprehension in language environments with limited resources.

5.3 Limitations

This study has a number of limitations in spite of its contributions:

- Dataset Imbalance: Model performance on minority categories is impacted by a notable class imbalance.
- Overfitting: The model exhibits significant overfitting, which restricts its capacity for generalization.
- Cultural Specificity: Limited use of humor and cultural allusions unique to a given area.
- Code-Mixing Complexity: Complex code-mixing patterns that go beyond basic Benglish are not fully addressed.
- Computational Requirements: The framework requires a significant amount of computing power for both inference and training.

5.4 Future Work

Several promising directions emerge for future research:

- Dataset Expansion: Creating more balanced datasets with enhanced representation of minority classes and different cultural references.
- Advanced Fusion Techniques: Exploring increasingly sophisticated fusion tactics, including cross-modal attention mechanisms and transformer-based fusion architectures. .
- More Classification: Though this research has only two class ..In future it can classify more categories
- Data Augmentation: Developing specialized augmentation algorithms for multimodal content to overcome class imbalance.
- Transfer Learning: Leveraging information from high-resource languages while adapting to Bengali-specific traits.
- Real-time Applications: Optimizing the framework for real-time classification in social media platforms and content control systems.

- Cultural Adaptation: Incorporating cultural and contextual knowledge to better perception of region-specific comedy and references.

5.5 Final Remarks

A major step toward comprehending multimodal digital content in low-resource languages like Bengali is represented by this research. Our work advances the larger objective of creating inclusive AI systems that can understand and interpret a variety of language and image expressions by tackling the particular difficulties of Bengali meme classification. In the end, this research contributes to more fair digital communication ecosystems by laying the groundwork for future developments in multimodal analysis for disadvantaged languages.

References

1. E. Hossain et al., “MemoSen: A Multimodal Bengali Meme Sentiment Analysis Dataset,” 2021. [Online]. Available: https: //github.com/eftekhar-hossain/Bengali-Hateful-Memes

2. V. Alluri et al., “Memotion Analysis: A Multimodal Benchmark Dataset for Meme Emotion Recognition,” 2020. [Online]. Avail- able: https://arxiv.org/abs/2112.11850

3. Parvin and M. M. Hoque, “An ensemble technique to classify multiclass textual emotion,” Procedia Computer Science, vol. 193, pp. 72–81, 2021, 10th International Young Scientists Conference in Computational Science, YSC2021, 28 June – 2 July, 2021. [Online]. Available: https://www.researchgate.

4.. P. Behera et al., “Multimodal Meme Sentiment Classification: SemEval-2020 Task 8,” in Proc. ICON, 2020. [Online]. Available: https://aclanthology.org/2020.icon-main.60

5. M. Islam et al., “SentimentFormer: Hybrid Transformer Fusion for Bengali Meme Sentiment Analysis,” 2023. [Online]. Avail- able: https://ieeexplore.ieee.org/document/103519946

6. F. Barbieri et al., “XLM-T: Multilingual Language Model for Social Media,” 2021. [Online]. Available: https://aclanthology. org/2022.lrec-1.27

7. V. Alluri and K. Krishna, “Deep Learning for Multimodal Meme Sentiment Analysis,” 2021. [Online]. Available: https://arxiv. Org/pdf/2112.11850

8. Jannat A. Das, O. Sharif, and M. M. Hoque, “An empirical framework for identifying sentiment from multimodal memes using fusion approach,” in Proceedings of 25th

International Conference on Computer and Information Technology (ICCIT). IEEE, 2022, pp. 791–796.

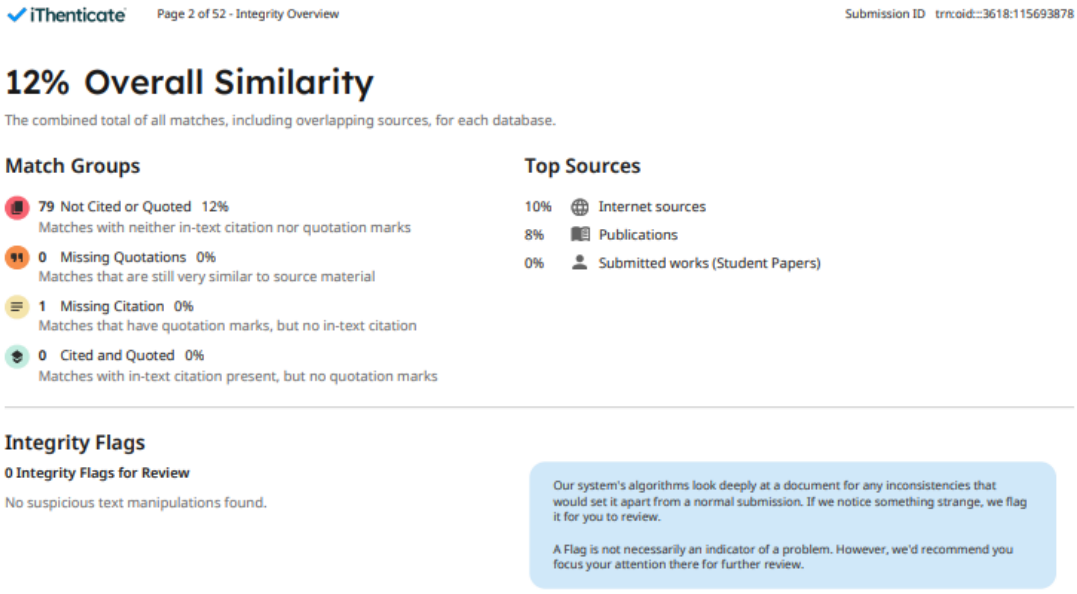
9. N. Elahi et al., “Explainable Multimodal Sentiment Analysis for Bengali Memes,” 2022. [Online]. Available: https://arxiv.org/ abs/2401.09446

10.Zou J. Ding and C. Wang, ”Utilizing BERT Intermediate Layers for Multimodal Sentiment Analysis,” 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 2022, pp. 1-6, doi. Available: 10.1109/ICME52920.2022. 9860014.

11. Taheri, A. C. Roy and A. Kabir, ”BEmoFusionNet: A Deep Learning Approach For Multimodal Emotion Classification in Bangla Social Media Posts,” 2023 26th International Con- ference on Computer and Information Technology (ICCIT), Cox’s Bazar, Bangladesh, 2023, pp. 1-6, Available: 10.1109/ ICCIT60459.2023.10441295.

12.Shanto, A. M., Priya, M. S. J., Tamim, F. S., & Hoque, M. M. (2025). MDC3: A Novel Multimodal Dataset for Commercial Content Classification in Bengali. NAACL Student Research Workshop. [Online] Available: https://aclanthology.org/2025.

Plagiarism Report



iThenticate Page 2 of 52 - Integrity Overview Submission ID trn:oid:::3618:115693878

12% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Match Groups

- **79** Not Cited or Quoted **12%**
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations **0%**
  Matches that are still very similar to source material
- **1** Missing Citation **0%**
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted **0%**
  Matches with in-text citation present, but no quotation marks

Top Sources

10% Internet sources
8% Publications
0% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.