

Article

Not peer-reviewed version

SentimentFormer: A Transformer-Based Multi-Modal Fusion Framework for Enhanced Sentiment Analysis of Memes in Under-Resourced Bangla Language

[Fatema Tuj Johora Faria](#) , [Laith H. Baniata](#) ^{*} , Mohammad H. Baniata , Mohannad A. Khair ,
[Ahmed Ibrahim Bani Ata](#) , [Chayut Bunterngrachit](#) , [Sangwoo Kang](#) ^{*}

Posted Date: 22 January 2025

doi: 10.20944/preprints202501.1587.v1

Keywords: early fusion; late fusion; intermediate fusion; bengali language; multimodal sentiment analysis; under-resource languages; social media; sentiment classification; machine learning






Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Article

SentimentFormer: A Transformer-Based Multi-Modal Fusion Framework for Enhanced Sentiment Analysis of Memes in Under-Resourced Bangla Language

Fatema Tuj Johora Faria ¹, Laith H. Baniata ^{2,*}, Mohammad H. Baniata³, Mohannad A. Khair⁴, Ahmed Ibrahim Bani Ata⁵, Chayut Bunterngrit⁶, and Sangwoo Kang ^{2,*}

¹ Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka 1208, Bangladesh

² School of Computing, Gachon University, Seongnam 13120, Republic of Korea

³ Computer Science Department, Faculty of Information Technology, The World Islamic Sciences and Education University, Amman 11947, Jordan

⁴ Qatrana Cement Company, Amman, Jordan

⁵ Department of Arabic Language, Faculty of Arts and Educational Sciences, Middle East University, Amman, Jordan

⁶ State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

* Correspondence: laith@gachon.ac.kr (L.H.B.); swkang@gachon.ac.kr (S.K.)

Abstract: Social media has increasingly relied on memes as a tool for expressing opinions, making meme sentiment analysis an emerging area of interest for researchers. While much of the research has focused on English-language memes, under-Resource languages, such as Bengali, have received limited attention. Given the surge in social media use, the need for sentiment analysis of memes in these languages has become critical. One of the primary challenges in this field is the lack of benchmark datasets, particularly in languages with fewer resources. To address this, we used the MemoSen dataset, designed for Bengali, which consists of 4,368 memes annotated with three sentiment labels: positive, negative, and neutral. MemoSen is divided into training (70%), test (20%), and validation (10%) sets, with an imbalanced class distribution: 1,349 memes in the positive class, 2,728 in the negative class, and 291 in the neutral class. Our approach leverages advanced deep learning techniques for multimodal sentiment analysis in Bengali, introducing three hybrid approaches. SentimentTextFormer is a text-based, fine-tuned model that utilizes state-of-the-art transformer architectures to accurately extract sentiment-related insights from Bengali text, capturing nuanced linguistic features. SentimentImageFormer is an image-based model that employs cutting-edge transformer-based techniques for precise sentiment classification through visual data. Lastly, SentimentFormer is a hybrid model that seamlessly integrates both text and image modalities using fusion strategies. Early Fusion combines textual and visual features at the input level, enabling the model to jointly learn from both modalities. Late Fusion merges the outputs of separate text and image models, preserving their individual strengths for the final prediction. Intermediate Fusion integrates textual and visual features at intermediate layers, refining their interactions during processing. These fusion strategies combine the strengths of both textual and visual data, enhancing sentiment analysis by exploiting complementary information from multiple sources. The performance of our models was evaluated using various accuracy metrics, with SentimentTextFormer achieving 73.31% accuracy and SentimentImageFormer attaining 64.72%. The hybrid model, SentimentFormer (SwiftFormer with mBERT), employing Intermediate Fusion, shows a notable improvement in accuracy, achieving 79.04%, outperforming SentimentTextFormer by 5.73% and SentimentImageFormer by 14.32%. Among the fusion strategies, SentimentFormer (SwiftFormer with mBERT) achieved the highest accuracy of 79.04%, highlighting the effectiveness of our fusion technique and the reliability of our multimodal framework in improving sentiment analysis accuracy across diverse modalities.

Keywords: early fusion; late fusion; intermediate fusion; bengali language; multimodal sentiment analysis; under-resource languages; social media; sentiment classification; machine learning

1. Introduction

The rapid growth of internet usage and the development of various Web 2.0 applications have led to a significant increase in the use of social media platforms such as Facebook, X (formerly known as Twitter), and Instagram. These platforms have transformed into spaces where users share their opinions on a wide range of topics, including business, politics, entertainment, and current events. Consequently, automated sentiment analysis of these conversations has attracted significant attention from Natural Language Processing (NLP) researchers, as it helps identify individuals' opinions or sentiments regarding specific events or issues. Most existing research in this field focuses on classifying textual sentiments into three primary categories: positive, negative, and neutral. [1,2]

However, the content shared on social media platforms is changing rapidly. More and more content is multimodal, combining images, text, and videos, which has added a new layer to sentiment analysis research [3]. Memes, for example, are becoming a popular way to share information. To understand the sentiment behind memes, it's important to consider multiple types of media at once. Since memes are often created in people's native languages, and social media usage is growing quickly in Bangladesh, there has been a rise in Bengali memes. Bengali, spoken by about 230 million people in Bangladesh and India, is one of the most spoken languages in the world [4].

This increase in Bengali memes has sparked more interest in sentiment analysis for different purposes. One key area is political sentiment analysis [5], which helps understand how people feel about policies and leaders. Other applications include social media emotion classification [6], which helps track user engagement and mental health, and detecting hate speech on online platforms to reduce harmful behaviors. Hate speech in Bengali is a serious concern as it covers a range of issues, from personal complaints to religious, political, and geopolitical conflicts [7,8]. Additionally, research on the level of toxicity against distinct groups in Bangla social media comments has highlighted the severity of harmful content [9]. Sentiment analysis is also being used in areas like news media, where it helps identify biases or frames in news articles and headlines [10]. These developments show how important and complex sentiment analysis is, especially when dealing with diverse types of content and languages.

The literature on sentiment analysis highlights key approaches for understanding emotional responses but also reveals significant gaps, particularly in the integration of multimodal data. Abiola et al. [11] conducted a study on emotional responses to the COVID-19 pandemic using sentiment analysis tools like TextBlob and VADER applied to tweets, uncovering the pandemic's impact on Nigeria's society, environment, and economy. However, while the study used topic modeling and visualized data, it lacked a multimodal approach, failing to incorporate visual data. Similarly, Sudirjo et al. [12] explored ChatGPT's potential in business customer sentiment analysis, emphasizing its ability to detect customer emotions. Yet, the study relied solely on text, missing the opportunity for multimodal sentiment analysis that could include images or videos. Faria et al. [5] examined political sentiment in the Bangladeshi elections, demonstrating the effectiveness of Pre-trained Language Models (PLMs) like BanglaBERT and Large Language Models (LLMs) like Gemini 1.5 Pro for sentiment detection. Despite the study's focus on Few-Shot learning, it did not utilize multimodal data to enhance sentiment analysis. Rifa et al. [13] proposed a sentiment analysis system for YouTube comments related to Bangla movies and dramas, introducing a dataset of 14,000 preprocessed comments for relevance detection and sentiment analysis. While the study made strides in analyzing sentiment using transformer models, it also lacked multimodal elements, relying solely on text. These studies, despite their contributions, share the limitation of not integrating visual data, which could have provided a more comprehensive understanding of sentiment by combining text with visual or sensory inputs. The absence of multimodal analysis in these works restricts their potential to capture the full range of human emotions and insights, underscoring the need for image-text pairing to deepen sentiment analysis capabilities.

To address the limitations of existing approaches in sentiment analysis for Bangla, this study introduces novel methodologies aimed at improving sentiment classification in this low-resource

language. We propose an integrated approach that combines unimodal text and image data with multimodal text-image pair analysis. By fine-tuning state-of-the-art pretrained models for both text and image data, we enhance the performance of sentiment detection. Furthermore, we explore various fusion strategies to effectively combine textual and visual information, improving accuracy and robustness in sentiment analysis. Through systematic hyperparameter tuning and rigorous evaluation using standard metrics, we ensure the models' optimal performance. Additionally, a comprehensive error analysis helps identify common misclassifications, providing valuable insights for future improvements in sentiment analysis for Bangla and other low-resource languages.

In this paper, we propose several hybrid methodologies aimed at improving sentiment classification for Bangla, a under-resourced language. The contributions of this study are summarized as follows:

- Proposed a three-fold approach for sentiment analysis in Bangla, incorporating unimodal text, unimodal image, and multimodal text-image pair data.
- Developed a systematic framework involving preprocessing, model development, and hyperparameter tuning for each modality, ensuring effective sentiment detection in Bangla.
- Fine-tuned state-of-the-art pretrained language models (mBERT, XLM-RoBERTa, DistilBERT) for Bangla sentiment classification, introducing a specialized framework tailored for text-based sentiment analysis in the Bangla language.
- Leveraged advanced image classification models (ViT, Swin Transformer, Swift Transformer) for sentiment analysis in images, and introduced a fine-tuned framework for enhancing visual sentiment detection.
- Introduced a hybrid framework combining both textual and visual modalities to improve sentiment classification accuracy, specifically addressing the challenges of sentiment analysis in Bangla.
- Explored three fusion strategies (Early Fusion, Late Fusion, Intermediate Fusion) to effectively combine text and image features, boosting performance in multimodal sentiment analysis for Bangla.
- Conducted systematic hyperparameter tuning for both text and image models, optimizing critical parameters to achieve the best possible performance while maintaining the models' ability to generalize.
- Provided a comprehensive evaluation using metrics such as accuracy, precision, recall, and weighted f1-score, offering valuable benchmarks for future research in sentiment analysis for the Bangla language.
- Performed a comprehensive error analysis for the multimodal approach to identify and address potential weaknesses in sentiment classification. This analysis examined both text and image modalities, pinpointing common misclassifications and their root causes, leading to insights for improving model performance and robustness.

The structure of this paper is as follows: Section 2 provides a comprehensive review of related literature, establishing the foundation for our research. Section 3 explores the relevant background studies. Section 4 describes the datasets utilized in this study. Section 5 outlines the proposed methodology in detail. Section 6 presents the experiments conducted and analyzes the results. Section 7 discusses the limitations of the study, Section 8 outlines potential directions for future research, and Section 9 summarizes the key findings and conclusions.

2. Literature Reviews

Sentiment analysis has seen substantial progress through both unimodal and multimodal approaches, with notable contributions leveraging diverse datasets and advanced machine learning techniques. Tables 1 and 2 summarize the key findings and methodologies from relevant studies in text-based and image-text pair-based sentiment analysis, respectively.

2.1. Unimodal (Text-Based) Approaches in Sentiment Analysis

Abiola et al. [11] analyzed emotional responses to COVID-19 by conducting sentiment analysis on over one million tweets from Nigeria, using TextBlob and VADER for sentiment classification and LDA for topic modeling. Their findings revealed that VADER classified 39.8% of the tweets as positive, 31.3% as neutral, and 28.9% as negative, while TextBlob identified 46.0% as neutral, 36.7% as positive, and 17.3% as negative. Despite the valuable insights provided by this study, it was limited by its unimodal approach, relying solely on text data. The incorporation of multimodal sentiment analysis, which could include images or videos, might have offered richer insights. Furthermore, the absence of transformer-based attention mechanisms in their methodology restricted the depth of sentiment interpretation, especially given that modern models like BERT were capable of offering more nuanced and context-aware sentiment analysis. Similarly, Manias et al. [14] explored multilingual approaches to sentiment and text classification in social media posts, focusing on BERT-based models and a zero-shot classification approach. Their study used four multilingual BERT models (mBERT cased, mBERT uncased, XLM-R, and DistilmBERT) to analyze multilingual datasets, finding that BERT-based classifiers excelled when fine-tuned on multilingual data, achieving high accuracy. While the zero-shot model was efficient and scalable, it provided relatively good results across multiple languages but lagged behind the fine-tuned models in terms of accuracy. The results demonstrated that XLM-R achieved an F1 score of 0.7642, showcasing its robust performance. However, similar to Abiola's study was also limited by its exclusive focus on text-based classification, without exploring multimodal approaches. Integrating multimodal data, such as images or videos, could have provided a more comprehensive understanding of social media content. Additionally, the reliance on pre-trained models without exploring domain-specific fine-tuning may have reduced the model's effectiveness for certain languages or tasks that were underrepresented in the training data. In contrast, Hu et al. [2] focused on sentiment analysis through advanced NLP techniques, such as ensemble methods, transfer learning, and deep learning architectures. By enhancing the robustness and precision of sentiment predictions, their approach investigated the impact of various models like recurrent neural networks and transformer-based architectures. They also introduced a novel ensemble method that combined multiple classifiers to improve predictive accuracy. However, like the previous studies, this research was limited to a text-based approach, focusing only on binary sentiment classification. Although the robustness of the models employed was notable, the absence of multimodal analysis in their study indicated an opportunity for more comprehensive sentiment analysis that integrated additional data types, such as images, videos, or audio. By incorporating multimodal data, future research could have provided a more nuanced understanding of sentiment in diverse social media contexts. In the same vein, He et al. [15] proposed a BERT-CNN-BiLSTM-Att hybrid model for sentiment analysis of short movie reviews, aiming to address challenges like polysemy and feature extraction in text sentiment analysis. The model employed BERT for dynamic word vectors, CNN for local feature extraction, and BiLSTM for global feature extraction, with an attention mechanism to highlight key information. Experimental results showed that the model outperformed alternatives like Word2Vec-BiLSTM and BERT-CNN, improving accuracy by up to 5.54%. However, like previous studies, this research was restricted to binary sentiment classification. Future research could have explored multiclass classification and expanded the dataset to include diverse elements, such as emoticons, which would have added to the richness of the analysis. Lastly, Gu et al. [16] predicted stock prices by integrating historical stock prices and financial news using the FinBERT-LSTM model. The methodology leveraged the pre-trained FinBERT for sentiment analysis of financial news and combined it with stock market data in an LSTM architecture to forecast stock prices. The results showed that the FinBERT-LSTM model outperformed both standalone LSTM and DNN models in prediction accuracy, as evidenced by metrics like Mean Absolute Error, Mean Absolute Percentage Error, and overall accuracy. The dataset used consisted of over 843,000 articles and stock price data spanning from 2009 to 2020. However, this study was limited by its reliance on only news sentiment and historical prices, potentially overlooking other influential factors that might have impacted stock

price predictions. In conclusion, while each of these studies contributed valuable insights to sentiment analysis, they all shared common limitations, such as their exclusive focus on text-based data and the absence of multimodal approaches. Incorporating multimodal data and exploring domain-specific fine-tuning could have enhanced the accuracy and depth of sentiment analysis across various domains.

Table 1. Summary of Studies on Unimodal (Text-Based) Sentiment Analysis.

Authors	Year	Models Employed	Performance Metrics	Key Findings
Abiola et al. [11]	(2023)	Sentiment analysis on 1M Nigerian tweets using TextBlob and VADER; LDA for topic modeling	VADER classified 39.8% positive, TextBlob identified 46.0% neutral, TextBlob was more accurate for neutral	Limited to text data; lacks transformer-based attention mechanisms for deeper sentiment interpretation
Manias et al. [14]	(2023)	Multilingual sentiment analysis using BERT-based models (mBERT, XLM-R, DistilmBERT)	XLM-R achieved F1 score of 0.7642, fine-tuned models performed better than zero-shot models	Limited to text-based classification; no multimodal analysis or domain-specific fine-tuning
Hu et al. [2]	(2024)	Sentiment analysis using ensemble methods, transfer learning, and deep learning (RNNs, transformers)	Naive Bayes (NB) achieved an F1 Score of 0.84	Focused on binary sentiment classification; no multimodal analysis
He et al. [15]	(2024)	Hybrid BERT-CNN-BiLSTM-Att model for sentiment analysis of short movie reviews	Improved accuracy by 5.54% compared to Word2Vec-BiLSTM and BERT-CNN	Restricted to binary sentiment classification; lacked diversity in dataset and multimodal elements
Gu et al. [16]	(2024)	FinBERT-LSTM model integrating stock prices and financial news for sentiment analysis	Fin-BERT Embedding LSTM Architecture achieved the highest accuracy of 0.955 at 77 epochs, outperforming other models	Relied solely on financial news and historical stock prices; overlooked multimodal data sources

2.2. Multimodal (Image-Text Pair-Based) Approaches in Sentiment Analysis

Elahi et al. [3] investigated the sentiment analysis of Bengali memes using the newly introduced MemoSen dataset, which fills a critical gap in low-resource language research. Specifically, their study combined ResNet50 for image processing and BanglishBERT for text analysis within a multimodal framework. Notably, this approach achieved a Weighted f1-score of 0.71, surpassing unimodal methods. Moreover, Explainable AI (XAI) was employed to interpret model behavior effectively. However, challenges such as an imbalanced dataset and relatively low accuracy were evident. Furthermore, a key limitation was the exclusive reliance on CNN-based architectures like ResNet50 and DenseNet161, without exploring Vision Transformer (ViT) models, which could have offered performance improvements. Similarly, Hossain et al. [1] introduced MemoSen, a novel Bengali multimodal dataset containing 4,368 memes annotated with sentiment labels (positive, negative, neutral). They also leveraged ResNet50 for visual analysis and Bangla-BERT for textual analysis. By utilizing early and late fusion techniques, their study achieved a Weighted f1-score of 0.643 and demonstrated a 1.2% improvement in multimodal sentiment classification over unimodal models. Nevertheless, like Elahi’s work did not incorporate Vision Transformers or modern variations for visual feature extraction, nor did it investigate intermediate fusion methods. These limitations highlight opportunities for further enhancements in model design and performance evaluation. On the other hand, Alluri et al. [17] focused on meme sentiment analysis using the Memotion dataset, which categorizes memes based on irony, humor, motivation, and overall sentiment. They exclusively employed Vision Transformers (ViT) for visual representation alongside advanced transformer-based models such as RoBERTa and

SBERT for textual and multimodal representations. Their multimodal approaches, including IMGTEXT, IMGSEN, and CAPSEN models, utilized fusion techniques to effectively integrate embeddings and achieved macro F1 scores of 0.633 for humor and 0.575 for overall sentiment. However, despite the robust use of transformer architectures and innovative fusion methods, their study was limited to English-language memes. Furthermore, they did not explore variations of Vision Transformer architectures, which could have provided diverse perspectives and potentially enhanced performance. In contrast, Thakkar et al. [18] addressed the gap in multimodal sentiment analysis by transforming a textual Twitter sentiment dataset into a multimodal format, emphasizing multilingual contexts. Their work utilized pre-trained models such as Multilingual-BERT, XLM-RoBERTa, CLIP, and DINOv2 for baseline experiments comparing unimodal and multimodal configurations. Through their pipeline, which integrated visual and textual features via concatenation followed by linear projection, they achieved strong results, particularly with sentiment-tuned large language models for text encoding. However, the study did not explore early, late, or intermediate fusion techniques, which could have provided deeper insights into feature integration and potentially improved classification accuracy. Taken together, these studies illustrate significant advancements in multimodal sentiment analysis, particularly for low-resource and multilingual contexts. However, common limitations, such as the lack of exploration into Vision Transformer architectures and intermediate fusion techniques, underscore the need for further investigation to enhance model performance and applicability across diverse datasets.

Table 2. Summary of Studies on Multimodal (Image-Text Pair-Based) Sentiment Analysis.

Authors	Year	Models ployed	Em-	Performance Metrics		Key Findings
Elahi et al. [3]	(2023)	ResNet50, EnglishBERT	Ban-	Weighted 0.71	f1-score:	Achieved higher performance than unimodal methods, utilized Explainable AI (XAI) to interpret model behavior. Limited by reliance on CNN architectures and absence of Vision Transformer (ViT).
Hossain et al. [1]	(2022)	ResNet50, Bangla-BERT		Weighted 0.643	f1-score:	Achieved 1.2% improvement in multimodal sentiment classification over unimodal models using early and late fusion techniques. Did not incorporate Vision Transformers or intermediate fusion methods.
Alluri et al. [17]	(2021)	Vision transformers (ViT), RoBERTa, SBERT	Trans- (ViT),	Macro F1 scores: 0.633 (humor), 0.575 (overall sentiment)		Utilized ViT for image processing and transformer-based models for text analysis. Limited to English-language memes and did not explore variations of Vision Transformer architectures.
Thakkar et al. [18]	(2024)	Multilingual-BERT, XLM-RoBERTa, CLIP, DINOv2		F1 Score: 76.8		Explored multimodal sentiment analysis for multilingual contexts, achieved strong results with sentiment-tuned large language models. Did not explore fusion techniques (early, late, intermediate).

3. Background Study

3.1. Models for Sentiment Analysis in Bangla Text

The rapid advancements in natural language processing (NLP) have revolutionized sentiment analysis, enabling robust and efficient classification of textual data. For Bangla, a low-resource and linguistically complex language, the development of state-of-the-art (SOTA) models has been particularly impactful. SOTA models such as multilingual BERT (mBERT), XLM-RoBERTa, and DistilBERT have set new benchmarks in Bangla sentiment analysis. These models excel in capturing

nuanced linguistic structures and sentiment expressions, making them indispensable tools for this task. By leveraging extensive pre-training on multilingual corpora and applying fine-tuning techniques to Bangla-specific datasets, these models achieve remarkable performance, even in resource-constrained scenarios.

3.1.1. mBERT (Multilingual BERT)

mBERT [22], or multilingual BERT, is an extension of Google's original BERT model, designed for multilingual NLP tasks. It employs multiple transformer encoder layers with self-attention mechanisms to understand complex relationships between words in different languages. Pre-trained on a diverse multilingual dataset using masked language modeling (MLM) and next sentence prediction (NSP) tasks, mBERT is highly versatile in cross-lingual tasks. For Bangla sentiment analysis, mBERT provides robust contextual understanding, effectively identifying nuanced sentiment expressions, even in the absence of large-scale annotated datasets.

3.1.2. XLM-RoBERTa

XLM-RoBERTa [23] is a multilingual variant of RoBERTa, optimized for cross-lingual tasks. Unlike mBERT, it focuses exclusively on MLM during pre-training, using massive multilingual corpora such as CommonCrawl to predict masked words. This focused training enhances its cross-lingual generalization and makes it particularly adept at low-resource languages like Bangla. XLM-RoBERTa has proven effective in sentiment analysis by capturing context-rich representations, identifying intricate sentiment cues in Bangla text, and simplifying multilingual workflows with automatic language detection.

3.1.3. DistilBERT

DistilBERT [24] is a lightweight and faster alternative to BERT, trained using knowledge distillation. It retains 97% of BERT's language understanding capabilities while being 40% smaller and 60% faster, making it an efficient option for sentiment analysis tasks. DistilBERT uses masked language modeling (MLM) as its primary pre-training objective. For Bangla sentiment analysis, DistilBERT is highly effective when computational resources are limited. Fine-tuning DistilBERT on Bangla sentiment datasets can yield competitive results, enabling the model to discern subtle sentiment patterns while maintaining high efficiency.

3.2. Models for Sentiment Analysis in Images

The rapid advancements in computer vision and multimodal learning have significantly transformed sentiment analysis in images, particularly in the domain of memes, where emotions, humor, and context are often conveyed visually. Effective sentiment analysis for memes requires not only the identification of visual elements but also an understanding of how these elements interact with text to express emotions. The ability to capture emotional cues from image features—such as facial expressions, body language, and scene composition—has become essential for accurately analyzing sentiment in memes. Recent models have leveraged sophisticated techniques, such as vision transformers and multimodal architectures, to enhance the analysis of emotions and sentiments from both the image and text components. Below, we explore several state-of-the-art models that have been developed to address these challenges.

3.2.1. Vision Transformer (ViT)

Vision Transformers (ViT) [19] provide a novel approach to image processing by using a transformer-based architecture rather than traditional Convolutional Neural Networks (CNNs). In the context of meme sentiment analysis, ViT divides images into fixed-size patches, which are then transformed into vector representations. These patches capture local image features, such as facial expressions or body language, which are crucial for detecting emotions. By incorporating positional encoding, ViT preserves the spatial relationships between image components, ensuring that important

features are properly contextualized. The model processes the sequence of patch embeddings through transformer encoder blocks, using self-attention mechanisms to understand how various parts of the image relate to one another. This enables ViT to capture global context in images, such as the interaction between the text and visual elements. The resulting image representations are classified for sentiment, making ViT highly effective for meme sentiment analysis.

3.2.2. Swin Transformer

The Swin Transformer [20] builds upon the Vision Transformer architecture by introducing a hierarchical approach to image processing. For meme sentiment analysis, this is particularly advantageous as it allows the model to capture both fine-grained local features (e.g., facial expressions) and broader global context (e.g., overall image composition). The image is divided into progressively smaller patches, which are processed at different hierarchical levels. This enables Swin Transformer to extract features across scales, enhancing its ability to capture complex emotional cues from both the image and its surrounding context. The shifted window-based self-attention mechanism ensures that the model focuses on important regions, such as the areas around faces or text, while maintaining global context. This hierarchical structure and attention mechanism make Swin Transformer well-suited for understanding the intricate relationships between visual and textual components in memes, allowing it to accurately predict sentiment.

3.2.3. SwiftFormer

SwiftFormer [21] introduces an efficient additive attention mechanism, which has been shown to reduce the computational complexity of traditional self-attention mechanisms. In the context of meme sentiment analysis, SwiftFormer can capture contextual information from images faster and with fewer resources. This is especially useful for real-time meme sentiment analysis on mobile devices or other resource-constrained environments. By using additive attention rather than matrix multiplication, SwiftFormer retains the ability to focus on important features within the image, such as emotional expressions or key text-image interactions, while significantly improving processing speed. This makes SwiftFormer an ideal choice for applications requiring fast and accurate meme sentiment analysis, even in environments with limited computational power.

3.3. Evaluation Metrics for Sentiment Analysis

In evaluating the performance of multimodal sentiment analysis tasks, several key metrics play crucial roles in assessing the effectiveness of the models:

3.3.1. Accuracy

Accuracy [25] serves as a fundamental metric for assessing the effectiveness of sentiment analysis models. It quantifies the proportion of correctly classified sentiment instances across both text and image modalities within the dataset. A higher accuracy score indicates that the model has successfully identified sentiment-related information from the multimodal data (e.g., text and image), demonstrating its ability to make correct predictions. Accuracy is a simple but essential metric in determining the model's overall performance in sentiment classification tasks.

$$\text{Accuracy} = \frac{\text{Number of correctly classified sentiment instances}}{\text{Total number of sentiment instances}} \quad (1)$$

3.3.2. Precision

Precision [26] in sentiment analysis refers to the proportion of instances that were correctly predicted as a specific sentiment (e.g., positive) out of all instances predicted as that sentiment. In the context of multimodal sentiment analysis in memes, precision measures how accurately the model identifies positive, negative, or neutral sentiments across all predicted instances of that sentiment.

For a specific sentiment class $c \in \{\text{positive, negative, neutral}\}$, precision is given by:

$$\text{Precision}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c} \quad (2)$$

Where:

- TP_c (True Positives for class c): The number of instances where sentiment c was correctly predicted as sentiment c .
- FP_c (False Positives for class c): The number of instances where sentiment c was incorrectly predicted, but the true sentiment was not c .

For the overall precision across all sentiment classes, we use the weighted average:

$$\text{Precision}_{\text{macro}} = \frac{1}{C} \sum_{c=1}^C \text{Precision}_c \quad (3)$$

where C is the number of sentiment classes (in this case, 3: positive, negative, neutral).

3.3.3. Recall

Recall [26] is the proportion of true instances of a specific sentiment class that were correctly identified by the model. In the context of sentiment analysis of memes, recall measures how well the model captures all instances of a specific sentiment, even if it results in false positives. Recall is critical when we aim to ensure that the model identifies every instance of a sentiment, such as detecting all positive or negative memes.

For a specific sentiment class $c \in \{\text{positive, negative, neutral}\}$, recall is given by:

$$\text{Recall}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c} \quad (4)$$

Where:

- TP_c (True Positives for class c): The number of instances where sentiment c was correctly predicted as sentiment c .
- FN_c (False Negatives for class c): The number of instances where sentiment c was incorrectly predicted as not c (i.e., the model missed an actual instance of sentiment c).

For the overall recall across all sentiment classes, we use the weighted average:

$$\text{Recall}_{\text{macro}} = \frac{1}{C} \sum_{c=1}^C \text{Recall}_c \quad (5)$$

where C is the number of sentiment classes (3 in this case: positive, negative, neutral).

3.3.4. Weighted f1-Score

The weighted f1-score [27] is an extension of the standard weighted f1-score that accounts for the class imbalances in a dataset by assigning different weights to different classes based on their frequency or importance. This metric provides a more accurate reflection of model performance when dealing with datasets where some sentiment classes (e.g., positive, negative, neutral) are underrepresented compared to others. In multimodal sentiment analysis, where the model is expected to analyze data from multiple modalities such as text and images, the weighted f1-score ensures that the evaluation is not disproportionately influenced by dominant classes. The weighted f1-score is calculated by averaging the weighted f1-scores for each class, with each weighted f1-score weighted by the support (the number of true instances) of that class. This allows for a more nuanced understanding of model performance, particularly in situations where some sentiment categories may be less frequent but still critical to the overall analysis.

$$F1_{\text{weighted}} = \sum_{i=1}^N w_i \cdot F1_{\text{Sentiment}_i} \quad (6)$$

Where:

- w_i is the weight for sentiment class i , calculated as:

$$w_i = \frac{\text{Number of true instances of sentiment class } i}{\text{Total number of instances}}$$

- $F1_{\text{Sentiment}_i}$ is the Weighted f1-score for sentiment class i , calculated as:

$$F1_{\text{Sentiment}_i} = \frac{2 \cdot P_{\text{Sentiment}_i} \cdot R_{\text{Sentiment}_i}}{P_{\text{Sentiment}_i} + R_{\text{Sentiment}_i}}$$

- $P_{\text{Sentiment}_i}$ is the precision for sentiment class i , calculated as:

$$P_{\text{Sentiment}_i} = \frac{\text{Correctly classified sentiment instances of class } i}{\text{Total predicted as sentiment instances of class } i}$$

- $R_{\text{Sentiment}_i}$ is the recall for sentiment class i , calculated as:

$$R_{\text{Sentiment}_i} = \frac{\text{Correctly classified sentiment instances of class } i}{\text{Total actual sentiment instances of class } i}$$

4. Dataset Description

In this study, we leverage the MemoSen [1] dataset, a multimodal dataset specifically curated for sentiment analysis in the Bengali language, to conduct our experiments. MemoSen was meticulously developed to address the lack of resources for multimodal sentiment analysis in Bengali. The dataset comprises 4,368 memes collected from popular social media platforms such as Facebook, Twitter, and Instagram over a period spanning February 2021 to September 2021. The memes were gathered using targeted keywords such as "Bengali Memes," "Bengali Funny Memes," and "Bengali Troll Memes," ensuring diverse representation across various themes. The dataset includes memes with captions written in Bengali, code-mixed (Bengali and English), or Banglish (code-switched). Memes failing to meet specific criteria, such as those lacking visual or textual components, containing unreadable text, or being duplicates, were excluded during curation. The final dataset is annotated into three sentiment categories: Positive, Negative, and Neutral, following rigorous guidelines to ensure consistency and reduce annotation bias. For training and evaluation purposes, the dataset is divided into train (70%), test (20%), and validation (10%) subsets. A detailed class-wise distribution is provided, along with representative examples of memes, including their captions and corresponding sentiment labels. The MemoSen dataset serves as a crucial benchmark for advancing research in multimodal sentiment analysis, especially for low-resource languages such as Bengali. Additionally, Figure 1 presents the distribution of samples across the training, test, and validation sets within the MemoSen dataset. Figure 2 showcases examples from the dataset, including the memes, their captions, and corresponding sentiment labels.

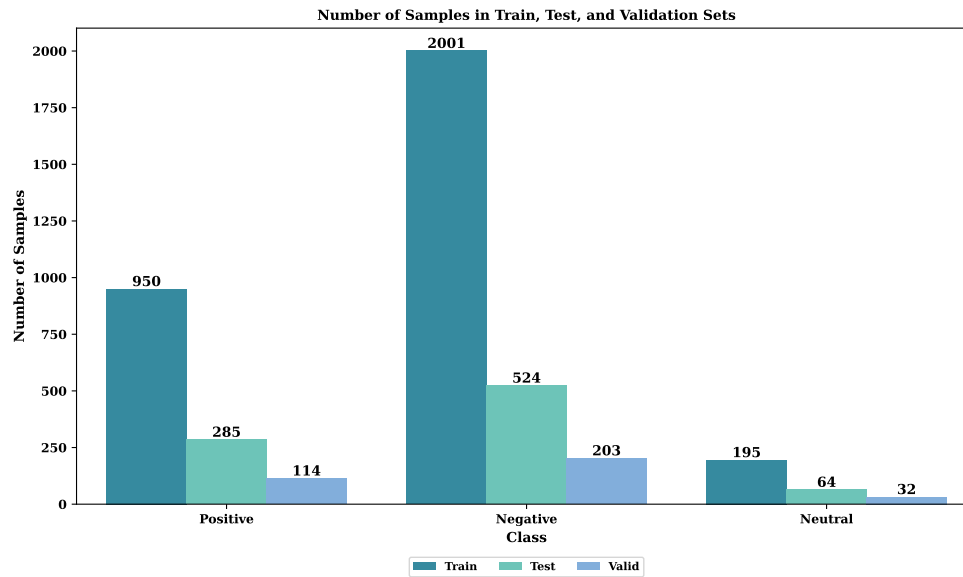


Figure 1. Distribution of Samples Across Train, Test, and Validation Sets in the MemoSen Dataset.

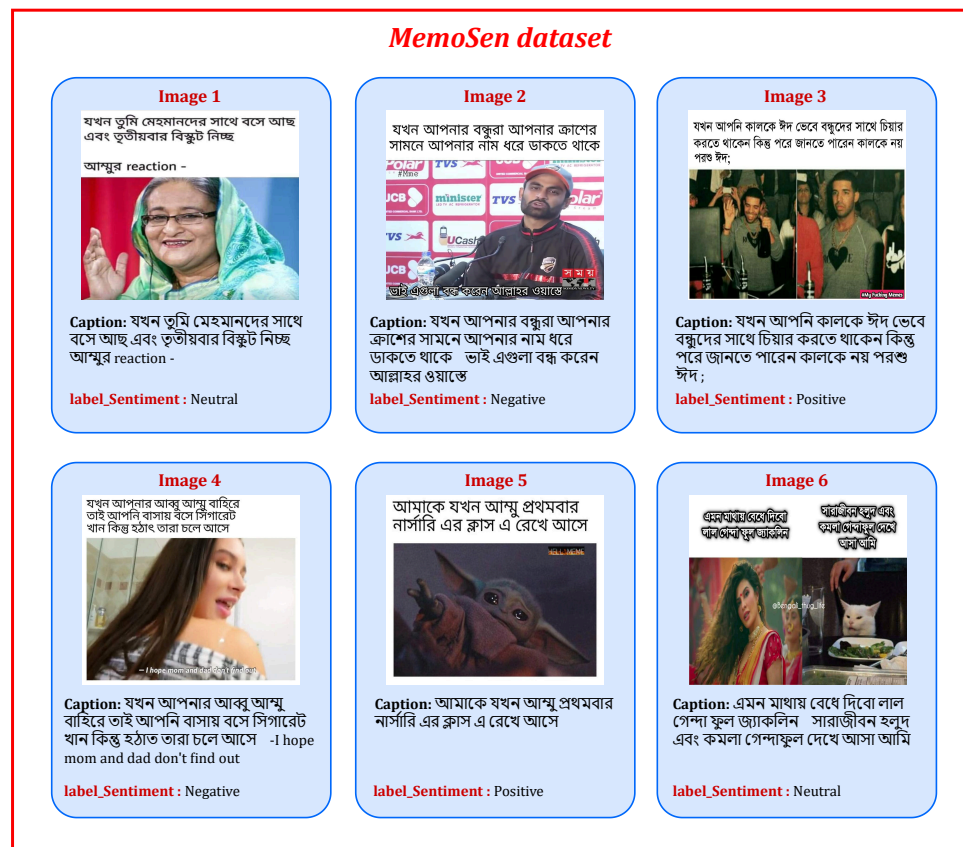


Figure 2. Representative examples from the MemoSen dataset, illustrating memes labeled with positive, neutral, and negative sentiments.

5. Proposed Methodology

We propose three hybrid approaches to multimodal sentiment analysis tailored for Bangla. "**SentimentTextFormer**" is a text-based method focused on accurately identifying sentiment-related information from Bangla texts. "**SentimentImageFormer**" introduces an image-based technique aimed at sentiment analysis, utilizing advanced transformer-based models for precise sentiment classification from visual data. Finally, "**SentimentFormer**" integrates text and image data through hybrid fusion techniques (Early Fusion, Late Fusion, and Intermediate Fusion), enhancing sentiment analysis capa-

bilities across multiple modalities in diverse contexts. Figures 3 and 4 highlight the architectures of SentimentTextFormer and SentimentImageFormer, respectively, while Figure 5 provides an illustrative overview of the SentimentFormer framework. All the code for these methodologies can be found on [GitHub](#).

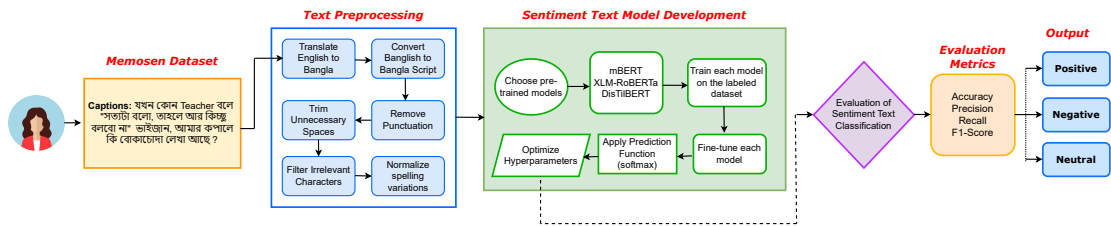


Figure 3. Unimodal Sentiment Classification Framework for Bangla Meme Captions.

5.1. Approach 1 for Unimodal Sentiment Analysis Framework for Bangla Captions

Step 1) Text Preprocessing: We preprocess Bangla sentiment data systematically to ensure consistency and relevance for analysis. First, we translate any English words or phrases into Bangla using Google Translator, ensuring the entire content remains in a single language. Next, we identify Banglish (Bangla written in Roman script) and convert it into proper Bangla script using tools like Google Translator or Gamista, maintaining linguistic uniformity. We then remove punctuation marks such as periods, commas, and exclamation points, as they often do not contribute to sentiment detection. Unnecessary spaces between words are eliminated to keep the text compact and avoid artificial lengthening. We also filter out irrelevant characters, including special symbols and control characters, that do not add to the semantic meaning of the sentiment. Lastly, we normalize spelling variations: In Bangla, there can be multiple ways to write the same word, especially with the influence of English or regional variations. We standardize these spelling differences to ensure consistency across the dataset. These steps ensure a clean, concise dataset that is optimized for accurate and efficient sentiment analysis.

Step 2) Sentiment Text Model Development: After text preprocessing, we move forward with developing the model for sentiment analysis. We utilize state-of-the-art pre-trained language models, including mBERT, XLM-RoBERTa, and DisTilBERT, all of which have proven effective in handling a variety of textual data. Each of these models was fine-tuned using the collected Bangla sentiment datasets to tailor them for the specific task of sentiment classification. Fine-tuning involves adjusting the model’s parameters using sentiment-labeled data to enhance its performance in categorizing Bangla text into different sentiment classes. This process helps the models better understand and classify the nuanced sentiments expressed in the Bangla language.

Step 3) Hyperparameter Tuning: For Bangla sentiment text identification, hyperparameter tuning plays a crucial role in optimizing model performance. This process involves fine-tuning key hyperparameters such as learning rate, batch size, dropout rate, and the number of training epochs, all of which significantly impact the model’s efficiency and performance. We systematically explore different configurations by adjusting these hyperparameters to find the optimal combination. Techniques such as grid search and random search are employed to automate this process, ensuring that the best-performing settings are identified. The model is trained multiple times with varying hyperparameter configurations, and the goal is to strike a balance between model complexity and generalization, allowing the model to accurately classify Bangla sentiment text while avoiding overfitting. Hyperparameter tuning helps in determining the best learning rate for the optimization process, the ideal batch size for training stability, and the appropriate dropout rate to prevent overfitting. Section 6.2 summarizes the results of hyperparameter tuning, showcasing the performance of each model under different settings.

Step 4) Evaluation of Sentiment Text Classification: After training and fine-tuning the models, we assess their performance in identifying sentiments from Bangla text. Each model is evaluated individually to determine its effectiveness in sentiment classification. The evaluation focuses on key performance metrics such as accuracy, precision, recall, and weighted f1-score. Accuracy measures the

overall correctness of the model in identifying sentiments, while precision reflects the model's ability to correctly identify positive sentiment instances. Recall, on the other hand, indicates how well the model identifies all the positive sentiment cases. The weighted f1-score accounts for class imbalances by combining precision and recall, providing a single, balanced measure of the model's performance across all sentiment categories. These metrics offer a comprehensive view of how well the models perform in classifying Bangla text into sentiment categories. Section 6.3 provides a detailed breakdown of the results, showing each model's performance across these key metrics.

5.1.1. Algorithmic Framework for Text-Based Bangla Sentiment Analysis

Text Preprocessing:

Given a Bangla text corpus $D = \{d_1, d_2, \dots, d_n\}$, the preprocessing begins by translating any English token $e \in d_i$ to Bangla b using:

$$T(e) = b, \quad \forall e \in d_i$$

where T is the translation function.

Next, punctuation and extraneous spaces are removed:

$$\hat{d}_i = \text{RemovePunct}(d_i), \quad \hat{d}_i = \text{TrimSpaces}(\hat{d}_i)$$

Irrelevant characters are then filtered out:

$$\tilde{d}_i = \text{FilterChars}(\hat{d}_i)$$

The cleaned dataset is represented as:

$$D' = \{\tilde{d}_1, \tilde{d}_2, \dots, \tilde{d}_n\}$$

Sentiment Text Model Development:

Let the fine-tuning function F operate on pre-trained models M and a labeled Bangla sentiment dataset L :

$$M^* = F(M, L)$$

For models $M \in \{\text{mBERT}, \text{XLM-RoBERTa}, \text{DisTilBERT}\}$:

$$M_j^* = F(M_j, L), \quad j = 1, 2, \dots, k$$

where k is the total number of models.

The output prediction function for sentiment classification is defined as:

$$P(d_i) = \text{softmax}(M^*(\tilde{d}_i))$$

Hyperparameter Tuning:

Let the hyperparameter space be H :

$$H = \{\eta, B, \lambda, E\}$$

where:

- η : Learning rate
- B : Batch size
- λ : Dropout rate
- E : Number of epochs

Define the performance function \mathcal{P} for a hyperparameter configuration $h \in H$:

$$\mathcal{P}(h) = \text{Evaluate}(M^*, h, L)$$

Optimization involves finding:

$$h^* = \arg \max_{h \in H} \mathcal{P}(h)$$

Search techniques can be applied as:

$$h^* = \begin{cases} \text{GridSearch}(H) & \text{if exhaustive search is feasible} \\ \text{RandomSearch}(H) & \text{otherwise} \end{cases}$$

Evaluation of Sentiment Text Classification:

Let the evaluation metrics include:

- **Accuracy (A):**

$$A = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision (P):**

$$P = \frac{TP}{TP + FP}$$

- **Recall (R):**

$$R = \frac{TP}{TP + FN}$$

- **Weighted f1-score (F1):**

$$F1 = 2 \cdot \frac{P \cdot R}{P + R}$$

For each model M_j^* , calculate:

$$\text{Metrics}_j = \{A_j, P_j, R_j, F1_j\}$$

5.2. Approach 2 for Unimodal Sentiment Analysis Framework for Meme Images

Step 1) Image Preprocessing: To ensure uniformity in the input size and optimize computational efficiency, we began by resizing the images to a consistent format of 224 x 224 pixels. This resizing step guarantees that the images are suitable for deep learning models while retaining important visual details. In addition, we utilized data augmentation techniques such as rotation, flipping, and other transformations to enrich the dataset and enhance the model's ability to generalize across various scenarios. These modifications introduce variations in the image orientations, which improves the model's robustness when handling new, unseen data. Furthermore, we performed image cleaning to eliminate any extraneous noise or artifacts that could interfere with the model's ability to detect sentiment-specific features. To optimize image quality, we applied image enhancement techniques, including adjustments to brightness, contrast, and sharpness, ensuring that the visual clarity of the images is maximized.

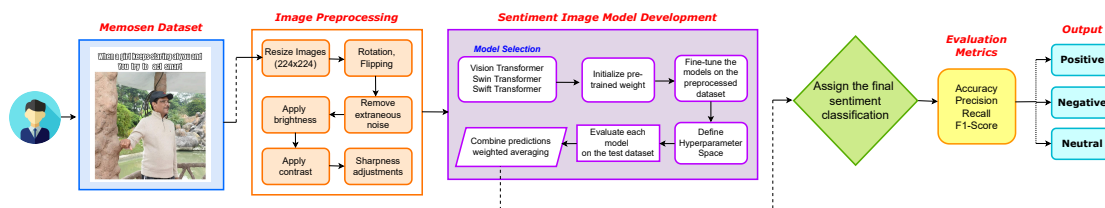


Figure 4. Unimodal Sentiment Classification Framework for Meme Images.

Step 2) Sentiment Identification Image Model Development: After preprocessing the images, we developed sentiment analysis models using three cutting-edge image classification models: ViT (Vision Transformer), Swin Transformer, and Swift Transformer. ViT leverages a transformer-based approach, dividing images into patches and processing them sequentially to capture global relationships and contextual information. Swin Transformer improves upon ViT by introducing a hierarchical structure and shifting window attention mechanism, enabling it to capture both local and global features more efficiently across varying image scales. Swift Transformer focuses on enhancing computational efficiency by simplifying the attention mechanism, ensuring faster processing without compromising accuracy in detecting sentiment features. By harnessing the unique strengths of these models, we aimed to accurately classify images based on their emotional or sentiment content.

Step 3) Hyperparameter Tuning: Hyperparameter tuning plays a crucial role in optimizing the performance of sentiment analysis models. This process involves adjusting several key parameters, such as the learning rate, batch size, and regularization strength, to find the optimal configuration that enhances model accuracy. The learning rate determines how quickly the model adjusts its weights during training, with higher rates speeding up learning but potentially causing instability, while lower rates offer slower, more stable convergence. Batch size influences how many images are processed before updating the model weights, with larger batches generally improving model stability but requiring more computational resources. Regularization strength helps prevent overfitting by penalizing complex models that may not generalize well to new data. During tuning, we evaluated model performance using metrics such as accuracy, precision, recall, and weighted f1-score, which collectively provide a comprehensive assessment of how well the model classifies sentiment in images. By systematically experimenting with different hyperparameter combinations, we identified the most effective settings for the models. Section 6.2 showcases the results of the hyperparameter tuning process, detailing the performance of each model under various configurations.

Step 4) Evaluation of Sentiment Image Analysis: After training and fine-tuning the models, we evaluated their ability to detect sentiment from images. Each model was assessed individually to determine its effectiveness in classifying sentiment. The evaluation process focused on essential performance metrics such as accuracy, precision, recall, and weighted f1-score. Accuracy measures the overall percentage of correct sentiment classifications, precision quantifies the model's ability to correctly identify positive sentiment instances, recall evaluates how well the model identifies all actual positive sentiment cases, and the weighted f1-score provides a balance between precision and recall. Section 6.3 offers a comprehensive analysis of the evaluation outcomes, highlighting the performance of each model based on these critical metrics.

5.2.1. Algorithmic Framework for Image-Based Bangla Sentiment Analysis

Image-Based Bangla Sentiment Analysis:

Given a dataset $\mathcal{D} = \{I_1, I_2, \dots, I_n\}$ containing n images, the preprocessing steps begin by resizing each image I_i to a uniform dimension of $a \times a$ pixels:

$$\hat{I}_i = \text{Resize}(I_i, a \times a), \quad \forall i \in \{1, \dots, n\}$$

Next, data augmentation techniques such as random rotation, flipping, and cropping are applied:

$$\tilde{I}_i = \text{Augment}(\hat{I}_i), \quad \forall i \in \{1, \dots, n\}$$

Noise removal and image enhancement (e.g., adjusting brightness, contrast, and sharpness) are then performed:

$$I'_i = \text{Enhance}(\tilde{I}_i), \quad \forall i \in \{1, \dots, n\}$$

Model Initialization:

Let the set of models be $\mathcal{M} = \{M_1, M_2, M_3\}$, where each model $M_j \in \mathcal{M}$ is initialized with pre-trained weights.

each model $M_j \in \mathcal{M}$:

$$M'_j = \text{FineTune}(M_j, \mathcal{D})$$

Hyperparameter Tuning:

Let the hyperparameter space be $H = \{\eta, B, D\}$, where:

- η : Learning rate
- B : Batch size
- D : Dropout rate

For each combination of η , B , and D , train and validate the model M_j :

$$M_j^* = \text{TrainValidate}(M'_j, H)$$

Record performance metrics such as accuracy, precision, recall, and Weighted f1-score for each configuration:

$$\mathcal{P}_j = \{A_j, P_j, R_j, F1_j\}, \quad \forall j \in \{1, 2, 3\}$$

Model Evaluation:

For each model M_j , evaluate it on the test dataset $\mathcal{D}_{\text{test}}$:

$$M_j^{\text{eval}} = \text{Evaluate}(M_j, \mathcal{D}_{\text{test}})$$

Compute evaluation metrics for each model:

$$\mathcal{P}_j = \{A_j, P_j, R_j, F1_j\}, \quad \forall j \in \{1, 2, 3\}$$

Prediction Aggregation:

Combine predictions from all models using majority voting or weighted averaging:

$$S_i = \arg \max_k \sum_j w_{M_j} \cdot P_{M_j}(S_i = k), \quad \forall i \in \{1, \dots, n\}$$

where w_{M_j} represents the model weights, and $P_{M_j}(S_i = k)$ is the probability of sentiment class k .

Final Sentiment Classification:

The final sentiment classification for each image is denoted as:

$$\mathcal{S} = \{S_1, S_2, \dots, S_n\}$$

where each S_i is the predicted sentiment for image I_i .

Return: Final sentiment classification \mathcal{S} .

5.3. Approach 3 for Exploring Different Fusion Techniques in Multimodal Sentiment Analysis

In our approach to multimodal Bangla sentiment analysis, we explored three different fusion techniques: Early Fusion, Late Fusion, and Intermediate Fusion.

Step 1) Feature Extraction: For text features, we leveraged advanced pre-trained language models, including mBERT, XLM-RoBERTa, and DistilBERT. These models are well-suited for extracting semantic and syntactic information from textual data, allowing us to effectively capture the nuances of sentiment-related expressions in Bangla. mBERT and XLM-RoBERTa excel at handling multilingual text, while DistilBERT provides a lighter and faster alternative while maintaining strong performance. From the text, we extract features such as word embeddings, contextual representations, and sentiment-specific tokens, which help in understanding the sentiment conveyed through the language. In parallel, we extracted image features using state-of-the-art models tailored for sentiment analysis tasks. These models include ViT, Swin Transformer, and Swift Transformer. Each model has been specifically trained to extract visual features from images, such as facial expressions, body language, color patterns, and

contextual visual elements, all of which are indicative of sentiment. ViT and Swin Transformer capture global image patterns, while Swift Transformer focuses on more localized, fine-grained image details. The extracted image features provide a rich representation of visual sentiment cues, complementing the text features for a more holistic analysis. By combining these distinct types of features—textual and visual—we ensure that both modalities contribute meaningfully to the sentiment analysis, enhancing the model's ability to accurately understand sentiment in a multimodal context.

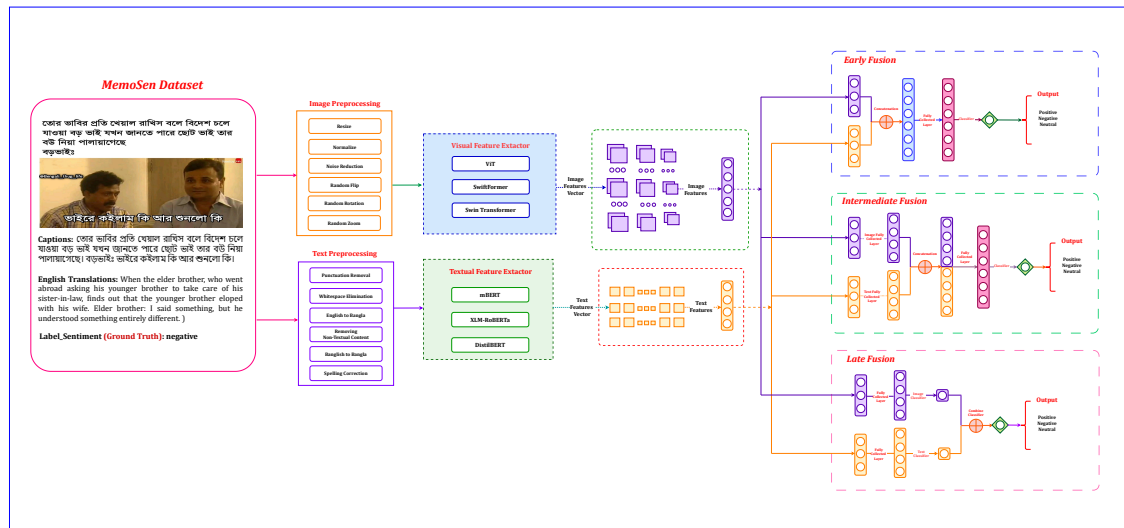


Figure 5. Fusion Framework for Enhanced Multimodal Sentiment Analysis of Bangla Memes.

Step 2) Fusion Techniques: We utilized three distinct fusion techniques to effectively combine textual and visual information in our Multimodal Bangla Sentiment Analysis pipeline. Prior to applying these techniques, we focused on extracting rich features from both text and images, ensuring that crucial information from each modality was thoroughly captured. These fusion strategies allow our model to take advantage of the complementary nature of text and image data, improving its ability to accurately analyze sentiment in a multimodal setting. By integrating both textual and visual insights, the model becomes more proficient at identifying and interpreting sentiment, considering both the language and the visual context.

- (a) **Early Fusion for Multimodal Sentiment Analysis:** For Early Fusion [28], we combine representations obtained from both text and image modalities at an early stage, prior to the sentiment classification process. This integration of features from multiple modalities facilitates the creation of joint representations, enabling a more nuanced understanding of sentiment by capturing both linguistic and visual cues. Let \mathbf{X} represent the input features from the text modality and \mathbf{Y} represent those from the image modality. The early fusion process can be mathematically described as:

$$\mathbf{z}_{\text{early}} = f_{\text{fusion}}([\phi_{\text{text}}(f_{\text{text}}(\mathbf{X})); \phi_{\text{image}}(f_{\text{image}}(\mathbf{Y}))])$$

In this equation:

- \mathbf{X} and \mathbf{Y} are the input features from the text and image modalities, respectively.
- $f_{\text{text}}(\cdot)$ and $f_{\text{image}}(\cdot)$ are the feature extraction functions for text and images, respectively.
- $\phi_{\text{text}}(\cdot)$ and $\phi_{\text{image}}(\cdot)$ are non-linear activation functions applied to the extracted features.
- $[\cdot; \cdot]$ denotes the concatenation operation, combining the features from both modalities.
- $f_{\text{fusion}}(\cdot)$ is the function that processes the concatenated features to produce the joint representation.
- $\mathbf{z}_{\text{early}}$ represents the fused features obtained from the early fusion process, which are then fed into a classifier to predict sentiment.

- (b) **Late Fusion for Multimodal Sentiment Analysis:** For Late Fusion [28], we aggregate predictions generated by text and image classification models at a later stage, after individual predictions are made. This technique integrates predictions from individual models to perform comprehensive sentiment analysis, potentially improving the accuracy and robustness of the final predictions. Let \mathbf{P}_{text} represent the prediction probabilities from the text sentiment classification model and $\mathbf{P}_{\text{image}}$ represent the prediction probabilities from the image sentiment classification model. The late fusion process can be represented as:

$$\mathbf{P}_{\text{fusion}} = \alpha \cdot \mathbf{P}_{\text{text}}(f_{\text{text}}(\mathbf{X})) + (1 - \alpha) \cdot \mathbf{P}_{\text{image}}(f_{\text{image}}(\mathbf{Y}))$$

In this equation:

- \mathbf{X} and \mathbf{Y} are the input features from the text and image modalities, respectively.
 - $f_{\text{text}}(\cdot)$ and $f_{\text{image}}(\cdot)$ are the feature extraction functions for text and images, respectively.
 - $\mathbf{P}_{\text{text}}(f_{\text{text}}(\mathbf{X}))$ represents the prediction probabilities from the text sentiment classification model applied to the text features.
 - $\mathbf{P}_{\text{image}}(f_{\text{image}}(\mathbf{Y}))$ represents the prediction probabilities from the image sentiment classification model applied to the image features.
 - α is a weighting factor that balances the contributions of text and image predictions, which can be fine-tuned for optimal performance.
 - $\mathbf{P}_{\text{fusion}}$ represents the final prediction probabilities obtained from the late fusion process, reflecting the overall sentiment classification result.
- (c) **Intermediate Fusion for Multimodal Sentiment Analysis:** For Intermediate Fusion [28], we merged features extracted from different modalities at an intermediate level of representation. By combining intermediate representations obtained from text and image processing pipelines, this technique captures the nuanced relationships between modalities, thereby facilitating more accurate sentiment analysis. Let \mathbf{Z}_{text} represent the intermediate features from the text modality, and $\mathbf{Z}_{\text{image}}$ represent the intermediate features from the image modality. The intermediate fusion process can be represented as:

$$\mathbf{Z}_{\text{fusion}} = f_{\text{fusion}}(\phi_{\text{text}}(f_{\text{text}}(\mathbf{X})), \phi_{\text{image}}(f_{\text{image}}(\mathbf{Y})))$$

In this equation:

- \mathbf{X} and \mathbf{Y} are the input features from the text and image modalities, respectively.
- $f_{\text{text}}(\cdot)$ and $f_{\text{image}}(\cdot)$ are the feature extraction functions for text and images, respectively.
- $\phi_{\text{text}}(\cdot)$ and $\phi_{\text{image}}(\cdot)$ are non-linear activation functions applied to the extracted features.
- $\mathbf{Z}_{\text{text}} = \phi_{\text{text}}(f_{\text{text}}(\mathbf{X}))$ represents the intermediate features from the text modality.
- $\mathbf{Z}_{\text{image}} = \phi_{\text{image}}(f_{\text{image}}(\mathbf{Y}))$ represents the intermediate features from the image modality.
- $f_{\text{fusion}}(\cdot, \cdot)$ is the fusion function that combines the intermediate features from both modalities.
- $\mathbf{Z}_{\text{fusion}}$ represents the fused features obtained from the intermediate fusion process.

Step 3) Hyperparameter Tuning: Hyperparameter tuning was performed to enhance the performance of the multimodal Bangla Sentiment Analysis models. This process involves adjusting several key parameters, including batch size, learning rate, fusion weight, and regularization strength, to determine the optimal configuration that maximizes performance metrics such as accuracy, precision, recall, and Weighted f1-score. Specifically, when applying fusion techniques to combine textual and visual information, hyperparameters play a crucial role in how these two modalities interact and contribute to the final sentiment prediction. In the context of fusion, parameters such as fusion weight determine how much influence the text features and image features will have in the final decision. For example, a higher fusion weight for text features may indicate that textual information is given more importance, while adjusting the weight for image features can help balance the contribution of visual cues. Other hyperparameters, such as the learning rate and batch size, help fine-tune how

quickly the model learns from the data and how much data it processes at once, directly impacting the efficiency and effectiveness of the fusion process. Furthermore, regularization parameters help prevent overfitting by controlling the complexity of the model, ensuring that the model generalizes well across unseen data. These tuning processes are essential for achieving the best performance from the multimodal model, as they allow the fusion techniques to adapt optimally to the specific characteristics of the Bangla sentiment analysis task. Section 6.2 display the results of hyperparameter tuning, showcasing how different configurations affect the model's performance across various fusion strategies. These tables provide a detailed comparison of how tuning different parameters influences the model's ability to analyze sentiment in both text and images.

Step 4) Evaluation of Multimodal Sentiment Analysis: The multimodal sentiment analysis models are evaluated to assess their performance in accurately identifying and classifying sentiment from both textual and visual data sources. Section 6.3 provide a detailed analysis of the performance metrics, such as accuracy, precision, recall, and weighted f1-score, which are computed to measure the effectiveness of the models. These metrics help assess how well the model integrates and interprets both textual and visual features for sentiment classification. Through rigorous evaluation, we ensure that the multimodal approach is robust and effective in real-world sentiment analysis tasks.

5.3.1. Algorithmic Framework for Multimodal-Based Bangla Sentiment Analysis

Feature Extraction: For text and image features, we use the following equations to express the extraction process.

For text feature extraction using pre-trained language models:

$$\mathbf{X} = f_{\text{text}}(\mathbf{T}), \quad \phi_{\text{text}}(\mathbf{X}) = \text{activation}(f_{\text{text}}(\mathbf{T}))$$

where:

- \mathbf{T} represents the raw text input,
- $f_{\text{text}}(\cdot)$ is the feature extraction function for text,
- $\phi_{\text{text}}(\cdot)$ is the activation function applied to the extracted features.

For image feature extraction using pre-trained models:

$$\mathbf{Y} = f_{\text{image}}(\mathbf{I}), \quad \phi_{\text{image}}(\mathbf{Y}) = \text{activation}(f_{\text{image}}(\mathbf{I}))$$

where:

- \mathbf{I} represents the raw image input,
- $f_{\text{image}}(\cdot)$ is the feature extraction function for images,
- $\phi_{\text{image}}(\cdot)$ is the activation function applied to the extracted features.

Fusion Techniques: We apply the following fusion strategies:

- Early Fusion:** The features from both text and image modalities are concatenated before classification. The fusion process is expressed as:

$$\mathbf{z}_{\text{early}} = f_{\text{fusion}}([\phi_{\text{text}}(f_{\text{text}}(\mathbf{X})); \phi_{\text{image}}(f_{\text{image}}(\mathbf{Y}))])$$

where:

- \mathbf{X}, \mathbf{Y} are text and image features,
 - $f_{\text{text}}(\cdot), f_{\text{image}}(\cdot)$ are feature extraction functions,
 - $\phi_{\text{text}}(\cdot), \phi_{\text{image}}(\cdot)$ are activation functions,
 - $f_{\text{fusion}}(\cdot)$ is the fusion function.
- Late Fusion:** The predictions from text and image classifiers are combined using a weighted sum:

$$\mathbf{P}_{\text{fusion}} = \alpha \cdot \mathbf{P}_{\text{text}} + (1 - \alpha) \cdot \mathbf{P}_{\text{image}}$$

where:

- $\mathbf{P}_{\text{text}}, \mathbf{P}_{\text{image}}$ are text and image classifier outputs,
- α is the fusion weight.

(c) **Intermediate Fusion:** The features from both modalities are fused at an intermediate stage:

$$\mathbf{Z}_{\text{fusion}} = f_{\text{fusion}}(\phi_{\text{text}}(f_{\text{text}}(\mathbf{X})), \phi_{\text{image}}(f_{\text{image}}(\mathbf{Y})))$$

where:

- \mathbf{X}, \mathbf{Y} are text and image features,
- $f_{\text{text}}(\cdot), f_{\text{image}}(\cdot)$ are feature extraction functions,
- $\phi_{\text{text}}(\cdot), \phi_{\text{image}}(\cdot)$ are activation functions,
- $f_{\text{fusion}}(\cdot, \cdot)$ is the fusion function.

Hyperparameter Tuning: Hyperparameters are tuned using the following expressions:
For the learning rate (η), batch size (B), and fusion weight (α):

$$\mathcal{L}_{\text{tune}} = \sum_{i=1}^N \mathcal{L}_{\text{loss}}(\mathbf{y}_i, \hat{\mathbf{y}}_i; \eta, B, \alpha)$$

where:

- $\mathcal{L}_{\text{loss}}(\cdot)$ is the loss function,
- $\mathbf{y}_i, \hat{\mathbf{y}}_i$ are the true and predicted labels,
- η is the learning rate,
- B is the batch size,
- α is the fusion weight.

The optimization process minimizes the loss $\mathcal{L}_{\text{tune}}$ to find the best hyperparameter settings.

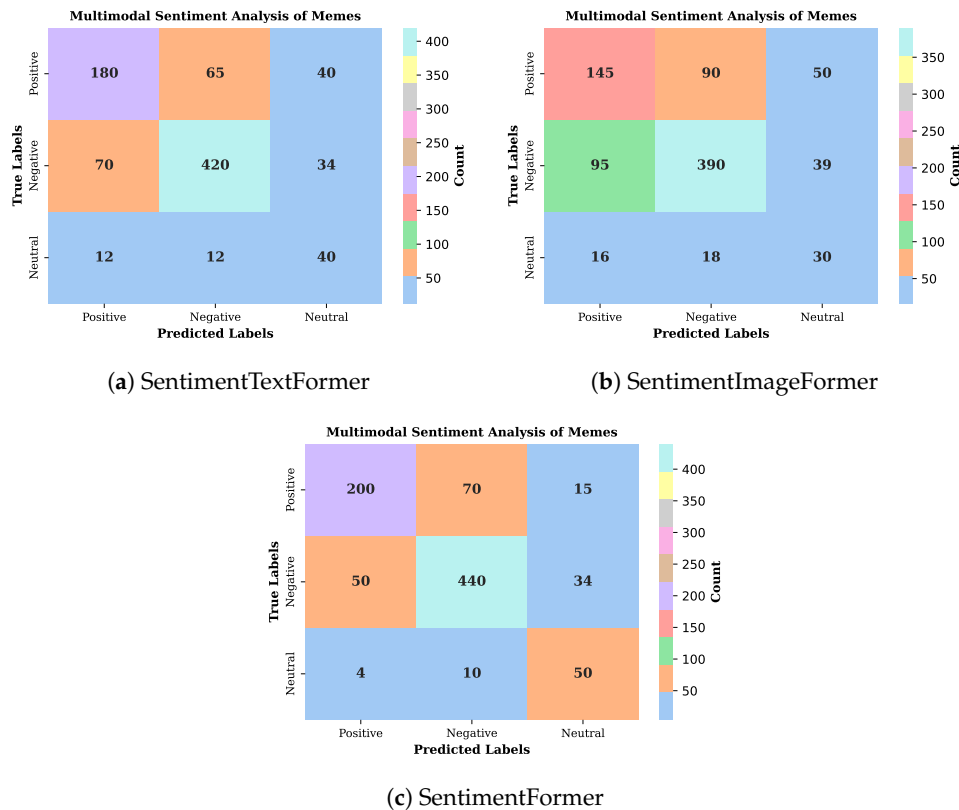


Figure 6. Confusion matrices of SentimentTextFormer, SentimentImageFormerPositive, and SentimentFormer showcasing their sentiment classification performance on the MemoSen dataset.

6. Experiments and Result Analysis

6.1. Experimental Setup

The experiments were conducted across four different environments: two Jupyter Notebooks, one Kaggle, and one Google Colaboratory. The first Jupyter Notebook setup featured an NVIDIA GeForce RTX 3050 GPU with a computing capability of 8.6, an Intel Core i5 9400f CPU, and 16 GB of RAM. The second Jupyter Notebook environment was equipped with a more advanced NVIDIA GeForce RTX 4060 Ti 16 GB GPU, an AMD Ryzen 7 5700X CPU, and 32 GB of RAM. Both Jupyter Notebook setups utilized Python 3.8.18 and PyTorch 2.0.1. The Kaggle environment provided access to Intel Xeon CPUs, 12.72 GB of RAM, and an NVIDIA Tesla P100 GPU with a compute capability of 6.0, running Python 3.10.13 with PyTorch 2.1.2. Lastly, the Google Colaboratory setup included a 15 GB Tesla T4 GPU, 12.5 GB of RAM, and PyTorch 2.3.1 with Python 3.10.12.

6.2. Hyper-Parameter Settings

The Table 3 provides the hyperparameter settings for both text-based and image-based models used in multimodal sentiment analysis of Bangla memes. For the image-based models, the Vision Transformer (ViT), Swin Transformer (SentimentImageFormer), and Swift Transformer are finetuned with a batch size of 8 and a learning rate of 1e-4, using the AdamW optimizer. The number of epochs varies slightly: ViT and Swift Transformer are finetuned for 45 epochs, while the Swin Transformer (SentimentImageFormer) is finetuned for 40 epochs. For the text-based models, mBERT (SentimentTextFormer), XLM-RoBERTa, and DistilBERT are finetuned with a batch size of 8, a learning rate of 1e-4, and the AdamW optimizer. These models are finetuned for either 45 or 50 epochs, with mBERT (SentimentTextFormer) being finetuned for 50 epochs and the others for 45 epochs.

Table 3. Hyperparameter Settings for Text-Based and Image-Based Models in Multimodal Sentiment Analysis of Bangla Memes.

Approch	Model	Batch Size	Epoch	Learning Rate	Optimizer
Image Based	ViT	8	45	1e-4	AdamW
	Swin Transformer (SentimentImageFormer)	8	40	1e-4	AdamW
	Swift Transformer	8	45	1e-4	AdamW
Text Based	mBERT (SentimentTextFormer)	8	50	1e-4	AdamW
	XLM-RoBERTa	8	45	1e-4	AdamW
	DistilBERT	8	45	1e-4	AdamW

The Table 4 provides the hyperparameter settings for early fusion models in the multimodal sentiment analysis of Bangla memes. For the early fusion models, the Vision Transformer (ViT) + mBERT, Swin Transformer + mBERT, and Swift Transformer + mBERT are all finetuned with a batch size of 8, a learning rate of 1e-4, and the AdamW optimizer. The number of epochs varies: ViT + mBERT is finetuned for 40 epochs, Swin Transformer + mBERT for 35 epochs, and Swift Transformer + mBERT for 30 epochs. For the ViT + XLM-RoBERTa, Swin Transformer + XLM-RoBERTa, and Swift Transformer + XLM-RoBERTa models, they are finetuned with a learning rate of 1e-3, a batch size of 8, and the AdamW optimizer, with epochs ranging from 35 to 40. Similarly, the ViT + DistilBERT, Swin Transformer + DistilBERT, and Swift Transformer + DistilBERT models are finetuned with a learning rate of 1e-4 and the AdamW optimizer, and the number of epochs is either 35 or 40 depending on the model.

Table 4. Hyperparameter Settings for Early Fusion in Multimodal Sentiment Analysis of Bangla Memes.

Approch	Model	Batch Size	Epoch	Learning Rate	Optimizer
Early Fusion	ViT + mBERT	8	40	1e-4	AdamW
	Swin Transformer + mBERT	8	35	1e-4	AdamW
	Swift Transformer + mBERT	8	30	1e-4	AdamW
	ViT + XLM-RoBERTa	8	35	1e-3	AdamW
	Swin Transformer + XLM-RoBERTa	8	40	1e-3	AdamW
	Swift Transformer + XLM-RoBERTa	8	35	1e-3	AdamW
	ViT + DistilBERT	8	35	1e-4	AdamW
	Swin Transformer + DistilBERT	8	35	1e-4	AdamW
	Swift Transformer + DistilBERT	8	40	1e-4	AdamW

The Table 5 provides the hyperparameter settings for late fusion models in the multimodal sentiment analysis of Bangla memes. For the late fusion models, the Vision Transformer (ViT) + mBERT, Swin Transformer + mBERT, and Swift Transformer + mBERT are all finetuned with a batch size of 8 and a learning rate of 1e-4, using the AdamW optimizer. The number of epochs varies: ViT + mBERT is finetuned for 40 epochs, Swin Transformer + mBERT for 35 epochs, and Swift Transformer + mBERT for 30 epochs. The models ViT + XLM-RoBERTa, Swin Transformer + XLM-RoBERTa, and Swift Transformer + XLM-RoBERTa are finetuned with a learning rate of 1e-3, a batch size of 8, and the AdamW optimizer, with epochs ranging from 35 to 40. Similarly, the ViT + DistilBERT, Swin Transformer + DistilBERT, and Swift Transformer + DistilBERT models are finetuned with a learning rate of 1e-4 and the AdamW optimizer, and the number of epochs is either 35 or 40 depending on the model.

Table 5. Hyperparameter Settings for Late Fusion in Multimodal Sentiment Analysis of Bangla Memes.

Approch	Model	Batch Size	Epoch	Learning Rate	Optimizer
Late Fusion	ViT + mBERT	8	40	1e-4	AdamW
	Swin Transformer + mBERT	8	35	1e-4	AdamW
	Swift Transformer + mBERT	8	30	1e-4	AdamW
	ViT + XLM-RoBERTa	8	35	1e-3	AdamW
	Swin Transformer + XLM-RoBERTa	8	40	1e-3	AdamW
	Swift Transformer + XLM-RoBERTa	8	35	1e-3	AdamW
	ViT + DistilBERT	8	35	1e-4	AdamW
	Swin Transformer + DistilBERT	8	35	1e-4	AdamW
	Swift Transformer + DistilBERT	8	40	1e-4	AdamW

The Table 6 outlines the hyperparameter settings for intermediate fusion models used in the multimodal sentiment analysis of Bangla memes. For the intermediate fusion models, the Vision Transformer (ViT) + mBERT, Swin Transformer + mBERT, and Swift Transformer + mBERT (SentimentFormer) are finetuned with a batch size of 8 and a learning rate of 1e-4 using the AdamW optimizer. The number of epochs varies slightly: ViT + mBERT is trained for 40 epochs, Swin Transformer + mBERT for 35 epochs, and Swift Transformer + mBERT for 30 epochs. The ViT + XLM-RoBERTa, Swin Transformer + XLM-RoBERTa, and Swift Transformer + XLM-RoBERTa models are finetuned with a learning rate of 1e-3, a batch size of 8, and the AdamW optimizer. These models are trained for 35 to 40 epochs. Similarly, the ViT + DistilBERT, Swin Transformer + DistilBERT, and Swift Transformer

+ DistilBERT models are finetuned with a learning rate of 1e-4 and the AdamW optimizer, with the number of epochs ranging from 35 to 40.

Table 6. Hyperparameter Settings for Intermediate Fusion in Multimodal Sentiment Analysis of Bangla Memes.

Approch	Model	Batch Size	Epoch	Learning Rate	Optimizer
Intermediate Fusion	ViT + mBERT	8	40	1e-4	AdamW
	Swin Transformer + mBERT	8	35	1e-4	AdamW
	Swift Transformer + mBERT (SentimentFormer)	8	30	1e-4	AdamW
	ViT + XLM-RoBERTa	8	35	1e-3	AdamW
	Swin Transformer + XLM-RoBERTa	8	40	1e-3	AdamW
	Swift Transformer + XLM-RoBERTa	8	35	1e-3	AdamW
	ViT + DistilBERT	8	35	1e-4	AdamW
	Swin Transformer + DistilBERT	8	35	1e-4	AdamW
	Swift Transformer + DistilBERT	8	40	1e-4	AdamW

6.3. Result Analysis

The Table 7 presents performance metrics for multimodal sentiment analysis of memes in Bangla, evaluated across Accuracy, Precision, Recall, and Weighted f1-score. Among text-based models, mBERT (SentimentTextFormer) leads with the highest accuracy (73.31%) and a Weighted f1-score of 64.34, followed by XLM-RoBERTa (72.85%, weighted f1-score 64.03) and DistilBERT (71.48%, weighted f1-score 62.29). For image-based models, ViT achieves the best accuracy (62.77%) but has lower precision and recall, resulting in an Weighted f1-score of 54.14. The Swin and Swift Transformers show similar performance, with accuracies of 64.72% and 63.57%, respectively.

Table 7. Performance Metrics of Text-Based and Image-Based Models for Multimodal Sentiment Analysis of Bangla Memes.

Approch	Model	Accuracy	Precision	Recall	Weighted f1-score
Text Based	mBERT (SentimentTextFormer)	73.31	62.77	68.60	64.34
	XLM-RoBERTa	72.85	62.38	68.35	64.03
	DistilBERT	71.48	60.9	66.14	62.29
Image Based	ViT	62.77	53.26	59.70	54.14
	Swin Transformer (SentimentImageFormer)	64.72	53.39	57.39	54.24
	Swift Transformer	63.57	53.90	59.84	54.79

The Table 8 presents performance metrics for multimodal-based models with early fusion in the context of sentiment analysis of Bangla memes. Among the model combinations, the Swin Transformer + XLM-RoBERTa achieves the highest accuracy (75.83%) along with solid precision (64.04%) and recall (67.68%), resulting in a weighted f1-score of 63.88%. The Swift Transformer + mBERT closely follows with an accuracy of 74.46%, precision of 63.24%, and recall of 68.82%, leading to a weighted f1-score of 63.69%. Another strong performer is the Swin Transformer + mBERT, which achieves an accuracy of 74.68%, with precision (62.97%), recall (67.04%), and a weighted f1-score of 63.03%. Other combinations, such as ViT + mBERT, ViT + XLM-RoBERTa, and ViT + DistilBERT, show lower performances, with accuracies ranging from 69.07% to 72.39%, and weighted f1-scores varying between 55.16% and 59.13%. These results demonstrate that early fusion of image-based models with text-based models, particularly

the Swin Transformer paired with XLM-RoBERTa, provides the best overall performance for Bangla meme sentiment analysis.

Table 8. Performance Metrics for Multimodal-Based Models with Early Fusion in for Multimodal Sentiment Analysis of Bangla Memes.

Approch	Model	Accuracy	Precision	Recall	Weighted f1-score
Early Fusion	ViT + mBERT	72.39	59.67	61.20	59.13
	Swin Transformer + mBERT	74.68	62.97	67.04	63.03
	Swift Transformer + mBERT	74.46	63.24	68.82	63.69
	ViT + XLM-RoBERTa	69.07	56.56	56.19	55.16
	Swin Transformer + XLM-RoBERTa	75.83	64.04	67.68	63.88
	Swift Transformer + XLM-RoBERTa	71.36	58.44	58.00	57.01
	ViT + DistilBERT	70.45	58.03	58.4	56.88
	Swin Transformer + DistilBERT	74.68	62.96	67.58	63.23
	Swift Transformer + DistilBERT	71.82	59.50	60.61	58.56

The Table 9 presents performance metrics for multimodal-based models with late fusion in the context of sentiment analysis of Bangla memes. Among the late fusion models, the Swin Transformer + XLM-RoBERTa achieves the highest accuracy (74.8%) with a precision of 60.38%, recall of 60.97%, and a weighted f1-score of 59.82%. The ViT + DistilBERT combination follows with an accuracy of 69.87%, precision of 55.69%, recall of 56.33%, and a Weighted f1-score of 55.28%. The Swift Transformer + DistilBERT also performs reasonably well, with an accuracy of 68.73% and a weighted f1-score of 54.68%. Other combinations such as ViT + mBERT, ViT + XLM-RoBERTa, and Swift Transformer + XLM-RoBERTa show lower performance, with accuracies ranging from 61.28% to 67.35%, and weighted f1-scores between 47.63% and 52.81%. These results demonstrate that late fusion models, particularly the Swin Transformer combined with XLM-RoBERTa, outperform other model combinations in terms of accuracy, precision, recall, and weighted f1-score for Bangla meme sentiment analysis.

Table 9. Performance Metrics for Multimodal-Based Models with Late Fusion in for Multimodal Sentiment Analysis of Bangla Memes.

Approch	Model	Accuracy	Precision	Recall	Weighted f1-score
Late Fusion	ViT + mBERT	61.28	48.78	48.05	47.63
	Swin Transformer + mBERT	71.02	56.6	56.97	56.09
	Swift Transformer + mBERT	67.35	53.93	54.78	52.81
	ViT + XLM-RoBERTa	62.43	49.66	48.84	48.49
	Swin Transformer + XLM-RoBERTa	74.8	60.38	60.97	59.82
	Swift Transformer + XLM-RoBERTa	62.77	49.61	50.1	49.02
	ViT + DistilBERT	69.87	55.69	56.33	55.28
	Swin Transformer + DistilBERT	65.29	52.94	53.25	51.98
	Swift Transformer + DistilBERT	68.73	55.04	56.23	54.68

The Table 10 presents performance metrics for multimodal-based models with intermediate fusion in the context of sentiment analysis of Bangla memes. Among the models with intermediate fusion, the Swift Transformer combined with mBERT (SentimentFormer) achieves the highest performance, with an accuracy of 79.04%, precision of 71.29%, recall of 77.42%, and a weighted f1-score of 73.28%.

Other notable models include the Swift Transformer + XLM-RoBERTa, which achieves an accuracy of 74.46%, precision of 65.12%, recall of 71.79%, and a weighted f1-score of 64.84%. The Swin Transformer + XLM-RoBERTa follows closely with an accuracy of 72.16%, precision of 62.85%, recall of 70.52%, and a weighted f1-score of 63.17%. In comparison, models such as ViT + mBERT and ViT + XLM-RoBERTa show lower performance, with accuracies ranging from 66.44% to 68.73%, and weighted f1-scores between 56.53% and 58.4%. These results indicate that intermediate fusion, particularly with the Swift Transformer and mBERT, leads to the best overall performance for Bangla meme sentiment analysis, outperforming other fusion strategies in terms of accuracy, precision, recall, and weighted f1-score.

Table 10. Performance Metrics for Multimodal-Based Models with Intermediate Fusion in for Multimodal Sentiment Analysis of Bangla Memes.

Approch	Model	Accuracy	Precision	Recall	Weighted f1-score
Intermediate Fusion	ViT + mBERT	68.16	57.83	63.45	58.4
	Swin Transformer + mBERT	68.73	59.8	68.08	60.43
	Swift Transformer + mBERT (SentimentFormer)	79.04	71.29	77.42	73.28
	ViT + XLM-RoBERTa	66.44	56.69	62.23	56.94
	Swin Transformer + XLM-RoBERTa	72.16	62.85	70.52	63.17
	Swift Transformer + XLM-RoBERTa	74.46	65.12	71.79	64.84
	ViT + DistilBERT	66.44	56.35	61.7	56.53
	Swin Transformer + DistilBERT	71.02	61.84	69.35	62.06
	Swift Transformer + DistilBERT	73.31	62.37	68.18	62.86

6.4. Error Analysis

In this section, we examine the limitations and misclassifications encountered during the multimodal sentiment analysis process in memes. By analyzing specific instances in which the model failed to accurately classify the sentiment (positive, negative, or neutral) in meme images and texts, we gain insights into the underlying challenges and areas for improvement. This analysis is crucial for understanding the model's weaknesses, such as difficulties in interpreting sarcasm, context, or visual cues, and for guiding the refinement of multimodal sentiment analysis capabilities. Figure 7 presents a visualization of error analysis for multimodal sentiment analysis in memes.



Figure 7. Error Analysis of Multimodal Sentiment Classification in Bengali Memes.

6.4.1. Misclassification of Humorous Memes Due to Lack of Contextual Understanding and Cultural Sensitivity in Sentiment Analysis

The image 1 is a meme featuring Tom and Jerry, with Tom sitting on a couch, reading a newspaper, and the text "I am not getting married now, I need my personal space for now" above him. Jerry, standing behind Tom, has the text "You will get married, even your father will get married" above him. Below the image, the text reads "desperate mother for marriage." The intended sentiment is likely positive, with humor derived from the contrast between Tom's desire for personal space and his mother's eagerness for him to marry, creating a relatable and exaggerated situation many young adults can understand. The humor lies in the exaggerated portrayal of the mother's determination. A model might misclassify it as negative due to several factors. First, text-based sentiment analysis might interpret phrases like "I am not getting married now" and "desperate mother" as negative indicators. Second, the model might lack contextual understanding, failing to grasp the humor in the situation. The contrast between Tom's desire for personal space and his mother's insistence on marriage is key to the humor, which might be missed. Additionally, cultural nuances could play a role, as a model might not fully understand the context of marriage in Bengali society, which the meme is referencing. Lastly, sarcasm and irony often found in memes can be challenging for models to detect, further contributing to the misclassification.

6.4.2. Misclassification of Social Awkwardness as Neutral Sentiment Due to Limited Contextual Understanding in Humorous Memes

The image 2 is a meme featuring a dialogue between two characters. The first character, a customer, is talking to the shopkeeper. The text above the customer reads: "Bought a condom, when I went to my girlfriend's house, I saw him there again." The text above the shopkeeper reads: "Uncle, you?" The intended sentiment of the meme is likely negative, as the humor arises from the awkward and embarrassing situation where the customer encounters the shopkeeper at his girlfriend's house. This unexpected encounter creates a humorous and relatable scenario, but the underlying situation is likely to cause embarrassment and discomfort for the customer. Several factors could lead a model to misclassify the sentiment as neutral: lack of contextual understanding, where the model might not grasp the social awkwardness and embarrassment implied in the situation; a focus on the literal meaning of the text, which doesn't explicitly convey negative emotions; limited training data, which might not cover similar scenarios involving social awkwardness and embarrassment; and challenges in detecting sarcasm or irony, which are often used in humor and can be difficult for models to interpret correctly.

6.4.3. Misclassification of Humor in Unexpected Interactions Due to Lack of Situational and Cultural Awareness in Sentiment Analysis

The image 3 is a meme featuring a dialogue between two characters. The text above the first character reads, "You won the big lottery, became a millionaire, didn't you?" and the text above the second character reads, "I am Jashim, are you?" The intended sentiment of the meme is likely neutral, with the humor stemming from the unexpected and seemingly random question posed by the second character. It creates a humorous disconnect between the first character's assumed wealth and the second character's seemingly irrelevant question. Several factors could lead a model to misclassify the sentiment as negative. First, the model might not grasp the humor in the situation, interpreting the unexpectedness of the question and the lack of a clear connection to the first character's wealth as dismissive or rude, which could lead to a negative sentiment classification. Additionally, the model might focus on the literal meaning of the text, which doesn't explicitly convey positive emotions, and fail to recognize the underlying humor and intended lightheartedness of the interaction. If the model was trained on a dataset lacking similar scenarios involving unexpected or random questions, it might struggle to classify the sentiment correctly. Cultural nuances could also play a role, as the humor might be lost on a model that lacks understanding of the Bengali language and the context of such interactions in Bengali society.

6.5. Comparison of Results with Existing Approaches

The Table 11 presents a comparison of the performance metrics—precision, recall, and Weighted f1-score—of three models for multimodal sentiment analysis of Bangla memes: the proposed SentimentFormer (Swift Transformer + mBERT), Hossain et al. (ResNet50 + CNN), and Elahi et al. (Banglish BERT + ResNet50). The SentimentFormer model outperforms both existing models in all metrics, achieving a precision of 71.29, recall of 77.42, and Weighted f1-score of 73.28. In comparison, the model by Hossain et al. scores 66.3 for precision, 62.8 for recall, and 64.3 for Weighted f1-score, while the model by Elahi et al. scores 69.0, 74.0, and 71.0, respectively. The SentimentFormer model shows significant improvements, particularly in recall (up by 14.62 over Hossain et al. and 3.42 over Elahi et al.) and Weighted f1-score (up by 8.98 over Hossain et al. and 2.28 over Elahi et al.), highlighting the effectiveness of combining Swift Transformer with mBERT and advanced multimodal fusion techniques. This demonstrates that the proposed method is more accurate and better at identifying true positive sentiment, making it a more balanced and robust approach for sentiment analysis in Bangla memes.

Table 11. Performance Comparison of Proposed Method with Existing Approaches for Multimodal Sentiment Analysis of Bangla Memes.

Model	Precision	Recall	Weighted f1-score
SentimentFormer (Proposed Method)	71.29	77.42	73.28
Hossain et al. [1] (ResNet50+CNN)	66.3	62.8	64.3
Elahi et al. [3] (Banglish BERT + ResNet50)	69.0	74.0	71.0

7. Limitations

The MemoSen dataset includes diverse multimodal data for Bangla sentiment analysis, sourced from publicly available platforms such as social media and news articles. While the dataset offers valuable insights for sentiment analysis research in the Bangla language, several inherent limitations must be considered when evaluating its comprehensiveness and generalizability. One key limitation is the representation of regional dialects within the Bengali language. Bengali is spoken in various regions, each with its own dialectal variations, yet the dataset primarily focuses on Bengali memes that may not adequately capture the full diversity of these regional dialects. As a result, the model may struggle to generalize across all Bengali-speaking communities, particularly those whose dialects are underrepresented or absent from the dataset. Another limitation is the restricted scope of sentiment categories. The dataset only includes three broad sentiment labels—Positive, Negative, and Neutral—which may fail to capture the full spectrum of emotions conveyed in memes. Memes often express nuanced sentiments such as sarcasm, humor, irony, or complex emotional gradients, which are difficult to encapsulate within these limited categories. Moreover, the temporal scope of the dataset, covering a specific time period from February to September 2021, introduces potential temporal biases. Meme culture and internet trends evolve quickly, and the sentiments expressed through memes may change over time. As such, the dataset may not fully represent current meme culture or the latest forms of sentiment expression in Bengali-language social media, limiting its applicability to more recent contexts. Additionally, the nature of memes often relies on humor, cultural references, and social commentary that may be rooted in stereotypes or specific societal contexts. As a result, the dataset may inadvertently reinforce or perpetuate negative stereotypes or biases, especially if certain types of memes are more likely to evoke specific sentiments based on cultural or social contexts. This could lead to a skewed understanding of sentiment within the dataset and affect the model’s performance in real-world applications where such biases are present.

8. Future Works

In future work, we plan to improve our approach to multimodal sentiment analysis for Bengali memes by exploring several exciting areas. One of the main improvements we want to make is using

Explainable AI (XAI) techniques, such as GradCAM++, LayerCAM, and ScoreCAM. These techniques will help us better understand how the model makes its predictions. They will show us which parts of the image and text are most important in deciding the sentiment. This transparency is important because it helps us understand the model's behavior, build trust in it, and make it work better. We also plan to use advanced Vision-Language Models (VLMs) like Claude 3.5 Sonnet and GPT-4, which excel at understanding and generating content that involves both images and text. By using these models, we aim to improve sentiment analysis in memes by generating responses based on different prompting techniques. These techniques could include providing specific instructions about the image or caption, asking the model to focus on certain emotions or elements, or even prompting it to consider various contextual cues. This approach will help the model capture subtle emotional clues, tones, and meanings in memes that simpler models might overlook. By refining the prompts, we can guide the model to generate more accurate and contextually aware responses, leading to a deeper understanding of sentiment in multimodal content. Additionally, we want to create a more inclusive and diverse dataset that includes different regional dialects of Bengali. This will involve collecting memes from areas like Chittagong, Sylhet, and Noakhali, which have unique dialects that are not often included in other datasets. Including these dialects will help our model work better for different Bengali-speaking communities and improve its performance in real-world situations. This will also ensure that the model understands the full richness of the Bengali language, including different cultural and regional expressions. By working on these areas, we hope to create more reliable, accurate, and understandable multimodal sentiment analysis models for Bengali. Our focus will be on capturing the different ways people express sentiments in regional languages and cultures.

9. Conclusions

In this study, we explored the emerging field of multimodal sentiment analysis for Bengali memes using the MemoSen dataset. This dataset consists of 4,368 Bengali memes annotated with sentiment labels (positive, negative, and neutral), offering a valuable resource for sentiment analysis in low-resource languages. By proposing and developing innovative hybrid models, SentimentTextFormer, SentimentImageFormer, and SentimentFormer, we demonstrated the potential of combining textual and visual information to improve sentiment classification accuracy. The use of advanced deep learning techniques, such as transformer-based models for both text and image modalities, along with fusion strategies like Early, Late, and Intermediate Fusion, significantly enhanced performance. Our models achieved notable results, with SentimentFormer (SwiftFormer with mBERT) reaching an accuracy of 79.04%, showing an improvement of 5.73% over the unimodal text model (SentimentTextFormer) and 14.32% over the unimodal image model (SentimentImageFormer). This demonstrates the effectiveness of our multimodal approach in outperforming both text-only and image-only models. However, there are some limitations in our work, such as the imbalanced class distribution in the MemoSen dataset, which could impact model performance, especially for the minority neutral class. Additionally, despite the improvements achieved, there is potential for further enhancement in handling more complex and diverse meme types. Future work will focus on addressing these limitations, including better handling of class imbalance, exploring more advanced fusion techniques, and expanding the dataset for greater generalization across different meme categories and sentiment nuances.

Author Contributions: F.T.J.F., L.H.B., M.H.B., M.A.K., A.I.B.A., C.B., and S.K. conceptualized and developed the methodology and experiments. F.T.J.F. conducted the experiments, while L.H.B. and S.K. analyzed the data. F.T.J.F. evaluated the results. F.T.J.F. wrote the manuscript, and L.H.B., M.A.K., A.I.B.A., and S.K. reviewed it. All authors have read and approved the final version of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Eftekhari Hossain, Omar Sharif, and Mohammed Moshikul Hoque. 2022. MemoSen: A Multimodal Dataset for Sentiment Analysis of Memes. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 1542–1554, Marseille, France. European Language Resources Association.
2. Zhengbing Hu, Ivan Dychka, Kateryna Potapova, Vasyl Meliukh, "Augmenting Sentiment Analysis Prediction in Binary Text Classification through Advanced Natural Language Processing Models and Classifiers", International Journal of Information Technology and Computer Science(IJITCS), Vol.16, No.2, pp.16-31, 2024. DOI:10.5815/ijitcs.2024.02.02
3. Elahi, K. T., Rahman, T. B., Shahriar, S., Sarker, S., Joy, S. K. S., & Shah, F. M. (2023, December). Explainable Multimodal Sentiment Analysis on Bengali Memes. In 2023 26th International Conference on Computer and Information Technology (ICCIT) (pp. 1-6). IEEE.
4. Faria, F. T. J., Moin, M. B., Wase, A. A., Ahmmed, M., Sani, M. R., & Muhammad, T. (2023). Vashantor: a large-scale multilingual benchmark dataset for automated translation of bangla regional dialects to bangla language. arXiv preprint arXiv:2311.11142.
5. Faria, F. T. J., Moin, M. B., Mumu, R. I., Abir, M. M. A., Alf, A. N., & Alam, M. S. (2024, September). Motamot: A Dataset for Revealing the Supremacy of Large Language Models over Transformer Models in Bengali Political Sentiment Analysis. In 2024 IEEE Region 10 Symposium (TENSYP) (pp. 1-8). IEEE.
6. A. Sharkar, M. A. R. Farhab, T. A. Tamanna, U. Rumman, M. T. R. Shawon and N. C. Mandal, "A Cross-Corpus Deep Learning Approach to Social Media Emotion Classification," 2022 4th International Conference on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh, 2022, pp. 1-6, doi: 10.1109/STI56238.2022.10103232.
7. Faria, F.T.J.; Baniata, L.H.; Kang, S. Investigating the Predominance of Large Language Models in Low-Resource Bangla Language over Transformer Models for Hate Speech Detection: A Comparative Analysis. Mathematics 2024, 12, 3687. <https://doi.org/10.3390/math12233687>
8. Karim, M. R., Dey, S. K., Islam, T., Shajalal, M., & Chakravarthi, B. R. (2022, November). Multimodal hate speech detection from bengali memes and texts. In International Conference on Speech and Language Technologies for Low-resource Languages (pp. 293-308). Cham: Springer International Publishing.
9. Moin, M. B., Debnath, P., Rifa, U. A., & Anis, R. B. (2024). Assessing the Level of Toxicity Against Distinct Groups in Bangla Social Media Comments: A Comprehensive Investigation. arXiv preprint arXiv:2409.17130.
10. Venugopal, J.P.; Subramanian, A.A.V.; Sundaram, G.; Rivera, M.; Wheeler, P. A Comprehensive Approach to Bias Mitigation for Sentiment Analysis of Social Media Data. Appl. Sci. 2024, 14, 11471. <https://doi.org/10.3390/app142311471>
11. Abiola, O., Abayomi-Alli, A., Tale, O.A. et al. Sentiment analysis of COVID-19 tweets from selected hashtags in Nigeria using VADER and Text Blob analyser. Journal of Electrical Systems and Inf Technol 10, 5 (2023). <https://doi.org/10.1186/s43067-023-00070-9>
12. Sudirjo, F., Diantoro, K., Al-Gasawneh, J. A., Khootimah Azzaakiyyah, H., & Almaududi Ausat, A. M. (2023). Application of ChatGPT in Improving Customer Sentiment Analysis for Businesses. Jurnal Teknologi Dan Sistem Informasi Bisnis, 5(3), 283-288. <https://doi.org/10.47233/jteksis.v5i3.871>
13. Rifa, Usafa & Debnath, Pronay & Kamal Rafa, Busra & Hridi, Shamaun & Rahman, Md. Aminur. (2024). CineXDrama: Relevance Detection and Sentiment Analysis of Bangla YouTube Comments on Movie-Drama using Transformers: Insights from Interpretability Tool. 10.48550/arXiv.2411.06548.
14. Manias, G., Mavrogiorgou, A., Kiourtis, A. et al. Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying twitter data. Neural Comput & Applic 35, 21415–21431 (2023). <https://doi.org/10.1007/s00521-023-08629-3>
15. A. He and M. Abisado, "Text Sentiment Analysis of Douban Film Short Comments Based on BERT-CNN-BiLSTM-Att Model," in IEEE Access, vol. 12, pp. 45229-45237, 2024, doi: 10.1109/ACCESS.2024.3381515.
16. Wen jun Gu, Yi hao Zhong, Shi zun Li, Chang song Wei, Li ting Dong, Zhuo yue Wang, and Chao Yan. 2024. Predicting Stock Prices with FinBERT-LSTM: Integrating News Sentiment Analysis. In Proceedings of the 2024 8th International Conference on Cloud and Big Data Computing (ICBDC '24). Association for Computing Machinery, New York, NY, USA, 67–72. <https://doi.org/10.1145/3694860.3694870>
17. SN. V. Alluri and N. Dheeraj Krishna, "Multi Modal Analysis of memes for Sentiment extraction," 2021 Sixth International Conference on Image Information Processing (ICIIP), Shimla, India, 2021, pp. 213-217, doi: 10.1109/ICIIP53038.2021.9702696.
18. Thakkar, G., Hakimov, S., & Tadić, M. (2024). M2SA: Multimodal and Multilingual Model for Sentiment Analysis of Tweets. arXiv preprint arXiv:2404.01753.

19. Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).
20. Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin transformer: Hierarchical vision transformer using shifted windows." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 10012-10022. 2021.
21. Shaker, A., Maaz, M., Rasheed, H., Khan, S., Yang, M. H., & Khan, F. S. (2023). Swiftformer: Efficient additive attention for transformer-based real-time mobile vision applications. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 17425-17436).
22. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
23. Conneau, Alexis, et al. "Unsupervised cross-lingual representation learning at scale." arXiv preprint arXiv:1911.02116 (2019).
24. Sanh, V. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.
25. Liu, Brian, and Madeleine Udell. "Impact of accuracy on model interpretations." arXiv preprint arXiv:2011.09903 (2020).
26. Goutte, C.; Gaussier, E. A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. In *Advances in Information Retrieval. ECIR 2005. Lecture Notes in Computer Science*; Losada, D.E., Fernández-Luna, J.M., Eds.; Springer: Berlin/Heidelberg, Germany, 2005; Volume 3408. https://doi.org/10.1007/978-3-540-31865-1_25
27. Sokolova, M.; Japkowicz, N.; Szpakowicz, S. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In *AI 2006: Advances in Artificial Intelligence, Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2006; Volume 4304, pp. 1015–1021. https://doi.org/10.1007/11941439_114.
28. Faria, F. T. J., Moin, M. B., Rahman, M. M., Shanto, M. M. A., Fahim, A. I., & Hoque, M. M. (2024). Uddesho: An Extensive Benchmark Dataset for Multimodal Author Intent Classification in Low-Resource Bangla Language. arXiv preprint arXiv:2409.09504.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.