MemeGuard: Transformer-Based Fusion for Multimodal Propaganda Detection in Low-Resource Social Media Memes

Md. Mohiuddin, Kawsar Ahmed
Shawly Ahsan and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u1904103, u1804017}@student.cuet.ac.bd
shawly.ahsan.bd@gmail.com, moshiul_240@cuet.ac.bd

Abstract

Memes are now a common means of communication on social media. Their humor and short format help messages spread quickly and easily. Propagandistic memes use both words and images to influence opinions and behaviors, often appealing to emotions or ideologies. While propaganda detection has been well-studied in high-resource languages (HRLs), there has been a limited focus on low-resource languages (LRLs), such as Bengali. In this study, we introduce **MemeGuard**, a new dataset of 3,745 memes for detecting propaganda in Bengali. We tested more than 45 different methods, including both single and combined approaches with fusion. For text, BanglaBERT-1 achieved the best macro F1 score of 80.34%, whereas the CLIP vision transformer scored 78.94% for images. The proposed multimodal model, which combines BanglaBERT-2 and CLIP via Adaptive Modality Fusion, achieved the highest macro-F1 score of 85.36%. This work establishes a strong baseline and offers valuable insights for future research in Bengali multimodal content analysis.

1 Introduction

Digital platforms have transformed human interaction by altering the ways individuals connect, share information, and express themselves. The proliferation of the Internet and Web 2.0 applications has fostered large, dynamic online communities, enabling rapid and accessible communication. Although digital openness offers significant advantages, it also accelerates the dissemination of misleading information, manipulative influence, and harmful narratives. On social media, memes function as a prominent and efficient communication medium, integrating concise text with impactful visuals to transmit messages (Zhong and Baghel, 2024). Certain memes are intentionally designed to manipulate audiences, influence opinions, and promote bias. These propagandistic memes advance specific political, religious, cultural, or ideological agendas by exploiting emotional responses or distributing misinformation (Cheng, 2025). As the influence of memes grows, identifying and analyzing propagandistic content has become essential. Because meme interpretation depends on both textual and visual elements, multimodal analysis is required for accurate detection and classification. While substantial research has addressed propaganda detection using text-based or multimodal methods in HRLs, LRLs such as Bengali remain underexplored. A comprehensive multimodal analysis for Bengali content is currently unavailable (Hossain et al., 2025). Detecting propagandistic memes in Bengali is necessary to support the dissemination of accurate information. However, no prior research has investigated propaganda detection in Bengali memes, leaving a significant research gap despite the growing prevalence of such content.

Building an automated system to detect propagandistic memes in Bengali presents several challenges. One major issue is the lack of a public dataset and the difficulty of extracting Bengali text from images, since there is no standard OCR tool for the language. Labelling memes by hand is also tricky because propaganda can be interpreted differently by different people. Memes often combine images and text, and the same image with different text can convey different meanings, adding complexity. Other problems include short text, discrepancies between the image and text, and the need to integrate both types of information. To address these problems, this study introduces a dataset of 3,745 Bengali-language memes for detecting propaganda. The study also proposes a transformer-based model that utilizes BanglaBERT-2 and CLIP, with an adaptive fusion of text and image features, to identify propaganda in memes more effectively. The main contributions of this work are:

- Developed **MemeGuard**, a multimodal dataset containing 3,745 memes, labelling propagandistic and non-propagandistic.
- Introduced a multimodal framework that combines textual and visual features using a late fusion strategy, where BanglaBERT-2 and CLIP models are employed with adaptive modality fusion to detect propaganda in memes effectively.

2 Related Work

Several studies have been conducted in various languages to detect propaganda in memes, including text, images, and multimodal content. This section provides a brief review of past studies on detecting memes, specifically propagandistic memes, across unimodal (e.g., text and image) and multimodal content.

2.1 Unimodal-based Propaganda Detection

Text-based propaganda detection has progressed considerably. Early work applied ML/DL with word embeddings. Noman et al. (2024) used a BiLSTM-CRF model for semantic web-based propaganda text, reporting F1 scores of 0.61 on multilingual and 0.688 on monolingual news data. Lichouri et al. (2023) examined disinformation detection using surface and morphological preprocessing, FastText vectors, and weighted TF-IDF fusion, obtaining a 77.60% F1-micro with LSVC, though effectiveness remained limited. A three-stage framework (Sourati et al., 2023) targeted logical fallacies in manipulative text. Building on such work, transformers and LLMs have been widely used for text classification, including propaganda detection. Ojo et al. (2023) conducted binary detection of persuasion strategies in Arabic news and tweets, achieving 64.00% F1 with XLM-RoBERTa. Horák et al. (2024) reported 92.26% F1 using XLM-RoBERTa Large for newspaper texts. Salman et al. (2023) found strong performance for code-switched English–Roman Urdu social media text using XLM-RoBERTa (Roman Urdu) and GPT-3.5-Turbo. Hasanain et al. (2024a) noted AraBERT outperforming GPT-4 for news articles, while Piña-García (2025) applied LLaMA 3.2 to political propaganda on Twitter.

In comparison, propaganda analysis using visuals alone has received far less attention than text-based approaches. Hs et al. (2021) used a DL-based ResNet-50 model and achieved 48.00% F1. More recently, Wang and Chen (2025) introduced a hybrid method for image-based propaganda detection. Koutlis et al. (2023) proposed Visual Part Utilization (VPU) with a ViT, reaching 94.98% accuracy but still excluding text. However, unimodal text- or image-only approaches fail to capture subtle context and often struggle with patriarchal content, underscoring the need for models that handle the complexities of multimedia content.

2.2 Multimodal-based Propaganda Detection

In addition to unimodal analysis, several multimodal approaches have been explored. Zaytoon et al. (2024) combined Bloomz-1b1 and ResNet101 with concatenation fusion for meme propaganda detection, achieving 80.51% F1-macro, though results were limited by an imbalanced Arabic dataset. Mahmoud and Nakov (2024) used VLM-generated descriptions with MPNet and CLIP-ViT for propaganda detection and multilabel classification, reporting 66.67% F1-macro but facing severe imbalance issues. Dimitrov et al. (2021) introduced a 950-meme corpus with 22 strategies, where VisualBERT COCO achieved 48.34% F1-micro, constrained by the small, imbalanced dataset. Alam et al. (2024a) created a 6,000-meme Arabic corpus with four classes, yielding weighted F1 scores of 69.00% (Qarib), 67.30% (ResNet50), and 65.90% (ConvNeXt, AraBERT, SVM), again limited by class imbalance. Qu et al. (2022) released Disinfomeme, a 1,170-meme dataset labelled as Disinfo or Non-Disinfo, where VisualBERT COCO achieved 53.3% on the BLM subset and 30.60% on the Veganism subset. Overall, these studies highlight the complexity of the task and suggest that improved fusion-based approaches may offer performance gains.

2.3 Multimodal Content Detection in Bengali

In contrast to other languages, propaganda detection using multimodal content in Bengali remains at a rudimentary stage. Existing multimodal studies using DL and transformer-based models have mainly addressed fake news detection (FAR, 2025), hate speech detection (Hossain et al., 2022), emotion classification (Rahman et al., 2025; Das et al., 2024), aggression detection (Hasan et al., 2025), and commercial content detection (Shanto et al., 2025). Multimodal content such as memes has also been used to detect aggressiveness (Alam et al., 2024b) and for sentiment analysis (Ahammad et al., 2025). Work leveraging LLMs for multimodal

classification is similarly limited. Hasan et al. (2024) examined LLMs with zero- and few-shot techniques for Bengali sentiment analysis and observed inferior performance with GPT-4. Building on this, Hossain et al. (2025) investigated VLMs, LLMs, MLMs, and vision transformers for multimodal text classification, showing that integrating pre-trained vision transformers for visual encoding and MLMs for textual encoding through fusion produced the best results.

Most existing studies in Bengali have primarily focused on detecting fake news, sentiment, aggression, and emotion, using mainly memes and text-image pairs. However, the detection of propaganda remains unexplored, and, to the best of our knowledge, there is currently no publicly available multimodal dataset specifically designed for propaganda detection, nor has a comprehensive multimodal analysis been conducted in this context. To address this gap, this study examines various multimodal techniques for the task on a newly developed dataset.

3 Dataset Development: MemeGuard

Our research reveals that no dataset currently exists for identifying multimodal propaganda in Bengali memes. To fill this gap, we developed **MemeGuard**, a multimodal dataset of 3,745 samples. This section describes the dataset's development process and key statistics.

3.1 Data Accumulation

The dataset was compiled by collecting Bengali memes over a five-month period from two sources: Facebook and a curated archive (Alam et al., 2024b). Facebook, a lively hub for Bengali user-created memes, provided dynamic content that captured local humour and cultural nuances. The archive offered structured collections that reflected diverse styles, in which textual variations could alter meaning. Keywords such as Bengali Memes, Bengali Funny Memes, Propaganda Memes, Bengali Celebrity Memes, Bengali Offensive Memes, and Bengali Political Memes were used to search groups and collect memes.

A total of 3,745 memes were collected, comprising 1,501 (40.1%) from Facebook and 2,244 (59.9%) from the curated archive (Fig. 1).

Only memes with Bengali captions were included, while excluding the following memes: (i) unimodal memes (text-only or image-only), (ii) memes containing unreadable text or very low quality image, and (iii) already existing memes. After that, the text was manually extracted from the browser using Google Lens, as Bengali lacks a reliable OCR system. Only relevant content was captured, excluding extraneous elements such as group names or creator identifiers. Finally, the extracted texts from memes were passed to annotators for manual labelling to ensure a rich dataset.

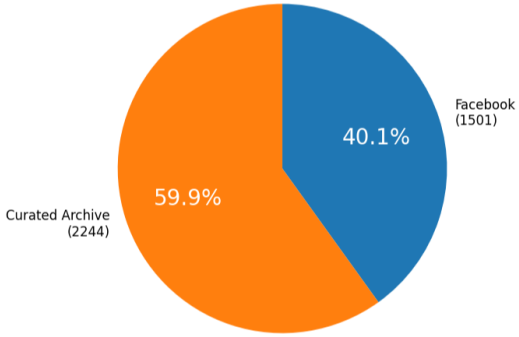


Figure 1: Distribution of data collection sources

3.2 Dataset Annotation

The MemeGuard dataset is designed for binary classification, categorizing memes into two distinct categories: Non-Propagandistic (Non-Prop) and Propagandistic (Prop). We follow the propaganda techniques proposed by Dimitrov et al., 2021 to define these categories in a simple, engaging, and precise way.

- **Non-Prop**: These memes are playful and neutral, designed to entertain without advancing any agenda. They often use humor, everyday scenarios, or light satire to foster connection through shared amusement.
- **Prop**: These memes seek to influence opinions or actions toward a specific goal. They promote political, ideological, or social agendas, often using emotionally charged or misleading content.

Guidelines for annotators are crucial for ensuring high-quality datasets. To assist annotators, we identified key characteristics as questions (see Appendix A), drawn from established propaganda techniques (Dimitrov et al., 2021), which are critical for distinguishing their manipulative nature. A team of five members conducted the manual annotation: four early-career NLP researchers (three graduate students and one research assistant) and

one senior NLP expert with 23 years of experience. The early-career annotators had 1–2.5 years of NLP experience, with 2 of them having prior annotation experience. Their ages ranged from 24 to 26, while the expert was 48.

A meme is considered propagandistic if it meets one or more characteristics defined in Appendix A. In the first stage, the three graduate annotators independently label the memes. Majority voting is applied to their labels to create the initial dataset. The RA annotator reviews this preliminary dataset. If the RA finds inconsistencies in labeling, the cases are discussed with the Expert Annotator to reach a final decision. This process produces the finalized dataset. We then calculated Cohen's kappa coefficient to measure inter-rater agreement (Cohen, 1960), with an average kappa value of 0.84. This shows nearly perfect agreement on the kappa scale, as shown in Table 8 (Appendix B), highlighting the robustness and dependability of the annotations for Bengali propagandistic meme detection.

3.3 Dataset Statistics

The MemeGuard dataset comprises 3,745 Bengali memes, with 865 labeled as *Prop* and 2,880 labeled as *Non-Prop*. The text contains 1,881 unique words, reflecting the linguistic diversity of Bengali memes. To support model training and evaluation, the dataset was stratified into 70% (2,621 memes) for training, 15% (562 memes) for validation, and 15% (562 memes) for testing, ensuring proportional representation of the 865 propagandistic and 2,880 non-propagandistic memes in each subset, as shown in Table 1. Specifically, the training set contains 605 propagandistic and 2,016 non-propagandistic memes, while the validation and testing sets each include 130 propagandistic and 432 non-propagandistic memes. Textually, memes average around 14 words per sample, with sentence lengths ranging from 2 to 66 words, and a total vocabulary exceeding 11,000 unique words. Visually, all images were resized to a uniform resolution of 224 × 224 pixels and stored in either PNG or JPEG format, with an overall average size of approximately 118 KB. This standardized and stratified setup ensures consistency, class balance, and reproducibility for robust model training and evaluation.

Figure 2 presents the frequency distribution of text lengths across the dataset, revealing that most memes contain between 5 and 20 words.

Class	**Train**	**Validation**	**Test**	*Total*
Prop	605	130	130	865
Non-Prop	2016	432	432	2880
Total	2621	562	562	3745
T_S	2621	562	562	3745
T_W	36340	8216	7834	52390
T_{UW}	9145	3444	3287	11308
L_{min}	2	3	2	-
L_{max}	55	66	47	-
$L_{avg.}$	13.93	14.73	13.96	-
I_T	2621	562	562	3745
$I_{avg.}$ (KB)	117.05	124.42	114.97	117.84
I_R (px)	**224 × 224**			
I_{format}	**PNG or JPEG**			

Table 1: Distribution of data across train, validation, and test sets. The symbols T_S, T_W, T_{UW} denote the total sentences, total words, and total unique words.

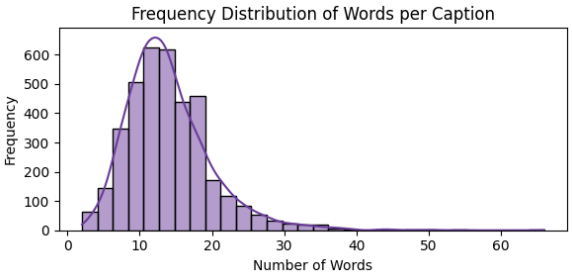


Figure 2: Frequency distribution of words per text

4 Methodology

A meme combines visual and textual elements, requiring parallel processing of both. This work examines deep learning and transformer-based models to extract and integrate these features. Feature-level fusion is then used to classify memes as propagandistic or non-propagandistic.

4.1 Data Preprocessing

Preprocessing standardizes inputs across modalities, thereby optimizing both learning and inference performance. Textual content preprocessing prepares data for classification models. Raw text is cleaned by removing stopwords, punctuation, and other unnecessary characters, reducing noise and improving feature extraction. The processed text is converted into dense vectors using embeddings like GloVe and FastText, or tokenized into IDs and attention masks with the Hugging Face tokenizer for transformer models such as BanglaBERT, MuRIL, and XLM-R. Padding adds extra tokens for uniform input lengths. Text normalization standardizes the format to ensure compatibility with pretrained vocabularies, especially for language-specific models like BanglaBERT.

For the visual modality, each image is resized to 224 × 224 pixels with three colour channels and converted to a tensor (a multidimensional array) using PyTorch, enabling GPU acceleration and batch processing. Normalization is performed using mean and standard deviation values from the ImageNet dataset, ensuring compatibility with the input requirements of standard image models such as VGG16, ResNet50, Swin Transformer, and Vision Transformer (ViT).

4.2 Unimodal Baselines

- **Text Modality:** Propagandistic meme detection explored various unimodal baselines that leverage deep learning and transformer-based architectures. For the text modality, CNN, BiLSTM, and CNN-BiLSTM hybrids were employed for binary classification, utilizing 300-dimensional GloVe and FastText embeddings. Training utilized the Adam optimizer, binary cross-entropy loss, and callbacks such as EarlyStopping and ReduceLROnPlateau. Transformer models included BanglaBERT-1 (Sarker, 2021), BanglaBERT-2 (Bhattacharjee et al., 2021), mBERT (Devlin et al., 2019), MuRIL (Khanuja et al., 2021), IndicBERT (Kakwani et al., 2020), BanglaElectra (NLP, 2024), and XLM-R (NLP, 2024). These pretrained Hugging Face models were fine-tuned on the developed corpus. Appendix C lists the tuned hyperparameters utilized for DL and transformer models for text modality.

- **Visual Modality:** The visual modality was addressed by sequentially applying pretrained convolutional neural networks (VGG16, VGG19, ResNet50, EfficientNet-B0, EfficientNet-B3 (Tan and Le, 2019)) and transformer-based models (ViT (Dosovitskiy et al., 2020), Swin (Liu et al., 2021), BEiT (Bao et al., 2021), DeiT (Touvron et al., 2021), ConvNeXT (Liu et al., 2022), CLIP (Radford et al., 2021)) for image feature extraction, each with varied fine-tuning strategies. Specifically, VGG19 (Simonyan and Zisserman, 2014) was fine-tuned on features [:15] using a custom classifier with linear layers, batch normalization, ReLU, and dropout. VGG16 was tuned up to features [:20] with fully connected layers, ReLU, and dropout. EfficientNetB3 unfroze the last 30 layers and employed global average pooling, batch normalization, dropout, and dense layers (ReLU, sigmoid). ResNet50 unfroze the previous 20 layers and used dropout and dense layers. EfficientNetB0 unfroze layers from 100 onward, with global average pooling, dropout, dense layers (ReLU), and sigmoid activation. For transformer-based models, pretrained versions were obtained from the Hugging Face collection and fine-tuned on the developed dataset. All models were trained with binary cross-entropy loss and class weights to address imbalance. Appendix C presents the various tuned hyperparameters used to create multiple visual models.

4.3 Multimodal Baselines

This work explores 16 multimodal baselines generated by combining the top-performing four textual (BanglaBERT-1, BanglaBERT-2, MuRIL, XLM-R) and four visual models (CLIP, BEiT, ViT, Swin), which merge their respective complementary strengths with various hyperparameters (Table 9 in Appendix C). The models were combined using feature-level fusion, where the [CLS] features from the textual and visual models were fused before generating logits, enabling effective integration of both modalities and enhancing the overall performance of the multimodal system.

4.3.1 Proposed Methodology

The proposed architecture for propagandistic meme detection in Bengali integrates two pre-trained models: BanglaBERT-2 for processing textual data and CLIP for analyzing image data. These models are combined using a late fusion approach, as illustrated in Figure 3, to leverage information from both modalities effectively. BanglaBERT-2 and CLIP models are fine-tuned on the developed dataset with manually tuned hyperparameters; both models are obtained from Hugging Face. To ensure consistency across modalities, both models were fine-tuned with carefully selected configurations derived through extensive manual experimentation. Let $V_{\text{logits}} \in \mathbb{R}^C$ and $T_{\text{logits}} \in \mathbb{R}^C$ denote the class logits produced by the visual and textual models, respectively, where C is the number of classes. The fusion procedure can be described as follows.

To calibrate the sharpness of each distribution, we apply temperature scaling with a learnable pa-

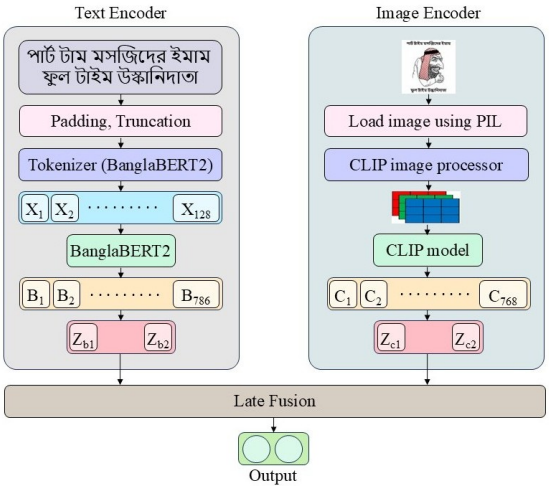


Figure 3: Architecture of the proposed multimodal approach for propaganda detection from memes.

rameter $\tau > 0$, with an initial value of 1.0 (Eq. 1).

$$V_{\text{scaled}} = \frac{V_{\text{logits}}}{\tau}, \qquad T_{\text{scaled}} = \frac{T_{\text{logits}}}{\tau}. \quad (1)$$

To adaptively control the relative contribution of each modality, we introduce learnable parameters $w_1, w_2 \in \mathbb{R}$ and compute modality weights via a softmax operation instead of giving equal weight as illustrated in Eq. 2.

$$[\alpha_1, \alpha_2] = \text{softmax}([w_1, w_2]), \quad \alpha_1 + \alpha_2 = 1, \quad \alpha_i \geq 0. \quad (2)$$

The fused logits are obtained as a convex combination of the scaled logits (Eq. 3.

$$\text{Final\_Logits} = \alpha_1 \cdot V_{\text{scaled}} + \alpha_2 \cdot T_{\text{scaled}} \quad (3)$$

Finally, the predicted label is chosen as the class with the largest value in the fused logits, as shown in Eq. 4.

$$\hat{y} = \arg \max_{c \in \{1,\ldots,C\}} \text{Final\_Logits}[c] \quad (4)$$

5 Experiments

The proposed framework was implemented and tested on a Kaggle GPU instance with a Tesla P100 GPU, 30 GB of RAM, and a Linux operating system supporting CUDA 11.8 and cuDNN. Python was used for development, utilizing PyTorch 2.1.0 and Hugging Face Transformers 4.35.2 for deep learning, as well as Pandas 2.0.3 and NumPy 1.24.3 for data processing. Experiments were run in Kaggle's Jupyter Notebook environment. Model performance was assessed using macro-F1, weighted-F1, and geometric mean (G-mean). All code and data are publicly available at https://github.com/MohiuddinPrantiq/MemeGuard-MultimodalPropagandaDetection.

5.1 Results and Analysis

Although the primary criterion for selecting the top-performing model was the macro F1-score (M-F1), which is suitable for imbalanced datasets, additional metrics, such as the weighted F1 score (W-F1) and G-Mean (G), were used for a comprehensive performance comparison.

5.1.1 Performance of Unimodal Baselines

Table 2 presents the performance of textual baselines. CNN+BiLSTM with FastText embeddings achieved superior performance (79.36&) over other DL models. Notably, Bangla BERT-1 achieved the highest M-F1 score of 80.34%, outperforming all other textual models tested in this study.

Model	M-F1	W-F1	G
CNN+GloVe	0.7069	0.7823	0.7212
CNN+ FastText	0.7069	0.7823	0.7212
BiLSTM+GloVe	0.6900	0.7600	0.7246
BiLSTM+FastText	0.7800	0.8400	0.7960
CNN+BiLSTN+ GloVe	0.7029	0.7717	0.7397
CNN+BiLSTN+FastText	**0.7936**	0.8491	0.8080
MuRIL	0.8011	0.8563	0.8070
BanglaBERT-2	0.7970	0.8549	0.7939
IndicBERT	0.7479	0.8159	0.7560
m-BERT	0.7879	0.8516	0.7666
XLM-R	0.8002	0.8550	0.8091
Bangla-Electra	0.7111	0.7793	0.7452
BanglaBERT-1	**0.8034**	0.8624	0.7826

Table 2: Performance of textual models.

Table 3 illustrates the performance of visual baselines. Among DL models, VGG19 and ResNet50 demonstrated strong performance, with VGG19 achieving the higher M-F1 score of 76.27%. However, CLIP achieved the highest macro F1-scores of 78.94%, demonstrating the superior representational capacity of transformer architectures in visual feature extraction for propagandistic meme detection.

5.1.2 Performance of Multimodal Models

Multimodal analysis, which utilizes both textual and visual modalities, enhanced performance across a feature-level fusion strategy, enabling effective integration and improving the overall performance of the multimodal system. As shown in Table 4, among the top four performing models per modality, **BanglaBERT-2+CLIP** achieved

Model	M-F1	W-F1	G
VGG16	0.5734	0.6219	0.6562
VGG19	**0.7627**	0.8438	0.6939
ResNet50	0.7500	0.8200	0.7386
EfficientNet-B0	0.4600	0.6800	0.1723
EfficientNet-B3	0.5500	0.7000	0.4473
ViT	0.7749	0.8400	0.7491
BEiT	0.7861	0.8534	0.7460
CLIP	**0.7894**	0.8587	0.7302
ConvNeXT	0.7384	0.8267	0.6642
DeiT	0.7643	0.8400	0.7367
Swin	0.7649	0.8400	0.7190

Table 3: Performance of visual models.

the highest M-F1 score of 82.83%. BanglaBERT-1+CLIP followed closely with an M-F1 score of 82.62%. Notably, CLIP consistently contributed to top-performing results across text models, highlighting its strong visual representation capabilities. The MuRIL – SWIN pair also demonstrated competitive performance, with a macro F1-score of 80.99%.

Text	Image	M-F1	W-F1	G
BanglaBERT-1	CLIP	0.8262	0.8785	0.8056
	BEiT	0.7970	0.8584	0.7733
	ViT	0.7837	0.8549	0.7243
	Swin	0.7979	0.8636	0.7444
MuRIL	CLIP	0.7673	0.8336	0.7614
	BEiT	0.7586	0.8339	0.7201
	ViT	0.7510	0.8277	0.7165
	Swin	0.8099	0.8630	0.8144
XLM-R	CLIP	0.8097	0.8664	0.7917
	BEiT	0.8016	0.8634	0.7661
	ViT	0.6728	0.7557	0.6848
	Swin	0.8014	0.8658	0.7494
BanglaBERT-2	**CLIP**	**0.8283**	**0.8802**	**0.8066**
	BEiT	0.8189	0.8741	0.7930
	ViT	0.8048	0.8655	0.7709
	Swin	0.7982	0.8588	0.7771

Table 4: Performance of multimodal combinations.

Following thorough hyperparameter tuning, the proposed model reached an M-F1 score of 85.36%. This represents a 5.02% improvement over the best text model (BanglaBERT-1, 80.34%), a 6.42% gain over the best visual model (CLIP, 78.94%), and a 7.02% increase compared to the best multimodal baseline (CLIP, 78.34%).

5.1.3 Impact of pre-trained multimodal baselines

Several prebuilt multimodal models were evaluated, including BLIP-2, CLIP, M-CLIP, and VisualBERT. Table 5 presents their performance metrics. However, these prebuilt solutions consistently underperformed relative to the custom fusion-based multimodal systems. Notably, CLIP achieved the highest M-F1 score among them at 78.34%, which remains significantly lower than that of the proposed method (BanglaBERT-2 + CLIP). This marked difference highlights the clear superiority of carefully designed fusion strategies over generic, end-to-end pre-trained multimodal models for classifying propagandistic memes in Bengali.

Model	M-F1	W-F1	G
BLIP-2	0.7300	0.8200	0.6425
CLIP	**0.7834**	0.8491	0.7577
M-CLIP	0.7430	0.8032	0.7838
VisualBERT	0.5755	0.6636	0.6034

Table 5: Performance of pre-trained multimodal models.

5.1.4 Impact of hyperparameters' tuning on performance

All models were trained using 70% of the dataset, validated on 15%, and tested on the remaining 15%. We conducted extensive hyperparameter tuning on the top-performing model (e.g., BanglaBERT-2+CLIP) to further enhance its performance. This process aimed to identify the optimal configuration to improve classification results and robustness. Table 6 provides a detailed overview of the selected hyperparameters and their impact on performance metrics. After tuning, with data split (80-10-10), late fusion, learning rate (5e-5), batch size (4), weight decay (0.1), and training for 20 epochs with gamma=2, the BanglaBERT-2+CLIP model achieved an M-F1 of 85.36%, about 2.53% higher than the initial configuration (82.83%), with corresponding improvements in W-F1 (89.86%) and G-Mean (82.97%), demonstrating that careful hyperparameter optimization significantly boosts performance.

Hyperparameter	Optimal	M-F1	W-F1	G
Data Split {60, 70, 80}	80-10-10	0.8324	0.8777	0.8523
Fusion Type {Feature, Late}	Late	0.8356	0.8843	0.8257
LR {(1,2,5) e-5, 5e-4}	5e-5	0.8406	0.8887	0.8236
Batch Size {4, 8, 16}	4	0.8536	0.8986	0.8297
WD {0.01, 0.1}	0.1	0.8536	0.8986	0.8297
Gamma & Epochs	2, 20	**0.8536**	**0.8986**	**0.8297**

Table 6: Performance across different hyperparameter configurations for BanglaBERT-2+CLIP model.

5.1.5 Comparison with existing techniques

To evaluate the effectiveness of the proposed model, we benchmarked its performance against several existing multimodal approaches (Zaytoon et al., 2024; Hasanain et al., 2024b; Qu et al., 2022) on the dataset we developed. Table 7 shows that the proposed method achieved the highest W-F1 score of 89.86%, surpassing all existing techniques and demonstrating superior capability in propagandistic meme detection across modalities. The proposed method outperforms the second-best approach (Bloomz-1b1+ ResNe101) by achieving an absolute improvement of 3.05% in M-F1 (from 85.36% to 88.41%) and 2.55% in W-F1 (from 89.86% to 92.41%).

Model	M -F1	W-F1	G
Bloomz-1b1 + ResNet101 (Zaytoon et al., 2024)	0.8231	0.8731	**0.8316**
ResNet + BERT to SVM (Hasanain et al., 2024b)	0.7293	0.8270	0.6243
VisualBERT-COCO (Qu et al., 2022)	0.6233	0.7016	0.6648
CLIP (Li et al., 2024)	0.7903	0.8509	0.7854
Proposed (BanglaBERT-2+CLIP)	**0.8536**	**0.8986**	0.8297

Table 7: Benchmarking of multimodal models on the test set.

5.1.6 Ablation Study

To assess the contributions of image-text modalities and fusion techniques, we perform an ablation study utilizing the macro F1 score. Text and image modalities perform competitively independently, with BanglaBERT-1 slightly outperforming CLIP (80.34% vs. 78.94%). Combining both using feature-level fusion—where representations from each modality are merged before the final classification—increases performance to 82.83%, demonstrating the complementary nature of the two approaches. The proposed late-fusion method integrates BanglaBERT-2 and CLIP by maintaining separate modality-specific representations and combining them only at the decision level, with improved hyperparameter tuning, which achieves the best macro F1 score of 85.36%. This represents gains of +2.53 points over intermediate fusion and +5.02 / +6.42 points compared to the text-only and image-only baselines, respectively, demonstrating that preserving modality-specific representations and integrating them later yields superior results for multimodal propagandistic meme detection.

5.2 Error Analysis

To gain an in-depth understanding of the proposed model's performance, a thorough error analysis is conducted using both quantitative and qualitative methods. The following parts present a detailed error analysis of the BanglaBERT-2+CLIP model.

5.2.1 Quantitative Error Analysis

The confusion matrix (Figure 4) confirms strong classification performance (337/374 correct; 90.11% accuracy). For the positive class, the error rate is 27.9% with 24 misclassified samples; for the negative class, the error rate is 4.5% with 13 misclassified samples, while the per-class weighted F1-scores are 0.7214 for *Non-prop* and 0.1771 for *Prop*, which sum to the overall weighted average of 0.8986. The lower weighted F1-score for propagandistic (0.1771) compared to non-propagandistic (0.7214) is primarily due to the severe class imbalance in the dataset, where class 1 accounts for only 23.1% (86 samples) of the total data. In comparison, class 0 dominates with 76.9% (288 samples). There are low false positives (4.5%) but higher false negatives (27.9%), indicating high precision yet room for improvement in recall. Overall, the results highlight the potential of multimodal transformer models for propagandistic content detection in multilingual contexts.

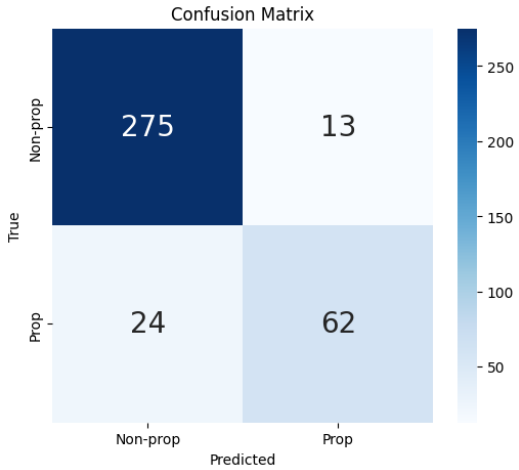


Figure 4: Confusion matrix of the proposed model.

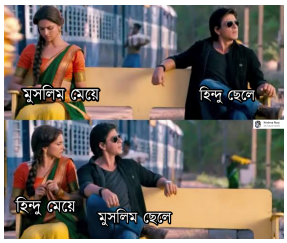
5.3 Qualitative Error Analysis

Figure 5 shows representative examples of correct and incorrect predictions generated by the proposed method (BanglaBERT-2+CLIP) compared with BanglaBERT-2 and CLIP. In Figure 5a, all three models correctly labelled the meme as non-

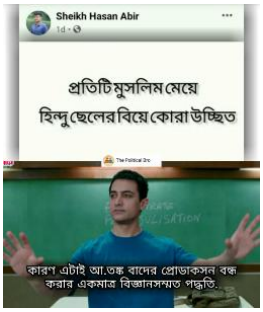
propagandistic, indicating strong agreement between visual and textual modalities when both cues align. In Figure 5b, the textual model misclassified the humorous content as propagandistic, and the multimodal model reflected this textual bias, whereas the visual model relied on image cues and correctly identified it as non-propagandistic, highlighting each model's dependency on its respective modality. Figure 5c shows a challenging case where the textual model detected propagandistic content, but the visual and multimodal models did not, due to reliance on visual cues that missed communal or cultural propaganda. The ground truth confirmed the propagandistic label, underscoring that models that prioritize visual context may overlook textual signals. Finally, Figure 5d shows that all modalities converged on the correct propagandistic classification, demonstrating that explicit textual propaganda, when supported visually, enables consistent predictions based on combined cues.

(a) Actual: Non-prop √
Textual: Non-prop √
Visual: Non-prop √
Multimodal: Non-prop √

(b) Actual: Non-prop
Textual: Prop X
Visual: Non-prop √
Multimodal: Prop X



(c) Actual: Prop
Textual: Prop √
Visual: Non-prop X
Multimodal: Non-prop X



(d) Actual: Prop √
Textual: Prop √
Visual: Prop √
Multimodal: Prop √

Figure 5: Examples of correct and incorrect predictions by the proposed model.

The four meme samples illustrate how each model's reliance on either textual, visual, or multimodal cues affects prediction accuracy. For instance, when both text and image suggest harmless humour, predictions are generally reliable across all models. However, challenges arise when propagandistic messages are covertly embedded in images that textual models may miss, or when culturally sensitive language in text prompts false positives from text-dependent models. These modality dependencies complicate the distinction between propaganda and non-propaganda, especially when nuanced religious or political content is present. Such subtleties can confuse models, leading to errors when one modality dominates interpretation. This dependency limits the models' ability to generalize across diverse meme formats, often leading to misclassification in emotionally or culturally complex instances and underscoring the importance of integrating and balancing multiple modalities for comprehensive understanding.

6 Conclusion

This work presents **MemeGuard**, a new dataset for detecting propaganda in Bengali memes. Using this dataset, forty-five unimodal and multimodal models are systematically evaluated for this task. Evaluation demonstrates that the BanglaBERT-2+CLIP model decisively surpasses all unimodal and multimodal baselines after fine-tuning on MemeGuard, achieving the top macro F1 score (85.36%) and weighted F1 (89.86%). These results highlight the strength of the proposed multimodal fusion in identifying propagandistic content. Future research will address current model limitations by expanding and diversifying the dataset, enhancing code-mixed data handling, exploring cutting-edge multimodal architectures such as LLMs and VLMs, refining fine-grained propaganda detection, and implementing automated hyperparameter optimization, while also leveraging a mix of our limited labeled corpus (3,745 samples) and additional unlabeled memes to enable semi-supervised learning, allowing the model to exploit abundant unlabeled data for richer representation learning—an approach particularly valuable for low-resource languages like Bengali. Beyond the well-established late-fusion strategy that yielded our best performance, future directions include investigating alternative fusion mechanisms, such as cross-modal attention, early or hierarchical fusion, and adaptive or gated fusion, as well as modality-specific architectural innovations, such as improved vision encoders, code-mixed language models, and joint embedding frameworks for more substantial multimodal alignment.

Limitations

Although the proposed method performs well at detecting propaganda in memes, several significant limitations remain unaddressed.

- A limited dataset size and narrow data sources may reduce generalizability.
- Class imbalance, with only 23.1% of samples labelled as propagandistic, may introduce bias in model training.
- Manual hyperparameter tuning is time-intensive and may not produce optimal results.
- Focusing solely on Bengali memes limits the method's applicability to other languages.

Ethics Statement

This study uses memes exclusively for academic research, without the intent to promote or distribute harmful material. Dataset creation and annotation adhered to copyright regulations, protected annotator privacy, and ensured fairness. Multiple annotators participated to minimize bias; however, cultural subjectivity persists as a limitation. The proposed system serves as a research instrument rather than a censorship mechanism, and human oversight remains necessary for practical implementation.

Declaration of AI Tools Uses

This manuscript complies with ACL policies on ethical AI use. ChatGPT (OpenAI) and Grammarly were used only for language refinement, such as improving grammar, clarity, coherence, and writing quality. The tools were not used to generate scientific ideas, analyze data, interpret results, or draw conclusions.

Acknowledgment

This work was supported by the Directorate of Research & Extension (DRE), Chittagong University of Engineering & Technology (CUET), Chittagong, Bangladesh, under the grant number: CUET/DRE/2023-2024/CSE/024.

References

2025. Multibanfakedetect: Integrating advanced fusion techniques for multimodal detection of bangla fake news in under-resourced contexts. *International Journal of Information Management Data Insights*, 5(2):100347.

Tanzin Ahammad, Shawly Ahsan, Jawad Hossain, and Mohammed Moshiul Hoque. 2025. M-sam: Multimodal sentiment analysis exploiting textual and visual features of social media memes. In *Pattern Recognition. ICPR 2024 International Workshops and Challenges*, pages 134–150, Cham. Springer Nature Switzerland.

Firoj Alam, Abul Hasnat, Fatema Ahmed, Md Arid Hasan, and Maram Hasanain. 2024a. Armeme: Propagandistic content in arabic memes. *arXiv preprint arXiv:2406.03916*.

Md Ashraful Alam, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024b. Multimodal aggressive meme classification using bidirectional encoder representations from transformers. In *2024 27th International Conference on Computer and Information Technology (ICCIT)*, pages 3542–3547. IEEE.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Kazi Samin, Md Saiful Islam, Anindya Iqbal, M Sohel Rahman, and Rifat Shahriyar. 2021. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. *arXiv preprint arXiv:2101.00204*.

Ziyang Cheng. 2025. Internet meme culture and political propaganda: The impact of 2024 us election memes on chinese online community. *Media Watch*, page 09760911251368965.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Avishek Das, Moumita Sen Sarma, Mohammed Moshiul Hoque, Nazmul Siddique, and M. Ali Akber Dewan. 2024. Avater: Fusing audio, visual, and textual modalities using cross-modal attention for emotion recognition. *Sensors*, 24(18).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting propaganda techniques in memes. *arXiv preprint arXiv:2109.08013*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Md Arid Hasan, Shudipta Das, Afiyat Anjum, Firoj Alam, Anika Anjum, Avijit Sarker, and Sheak Rashed Haider Noori. 2024. Zero-and few-shot prompting with llms: A comparative study with fine-tuned models for bangla sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17808–17818.

Md. Maruf Hasan, Shawly Ahsan, Mohammed Moshiul Hoque, and M. Ali Akber Dewan. 2025. Mulad: Multimodal aggression detection from social media memes exploiting visual and textual features. In *Pattern Recognition*, pages 107–123, Cham. Springer Nature Switzerland.

Maram Hasanain, Fatema Ahmad, and Firoj Alam. 2024a. Can GPT-4 identify propaganda? annotation and detection of propaganda spans in news articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2724–2744, Torino, Italia. ELRA and ICCL.

Maram Hasanain, Md Arid Hasan, Fatema Ahmed, Reem Suwaileh, Md Rafiul Biswas, Wajdi Zaghouani, and Firoj Alam. 2024b. Araieval shared task: propagandistic techniques detection in unimodal and multimodal arabic content. *arXiv preprint arXiv:2407.04247*.

Aleš Horák, Radoslav Sabol, Ondřej Herman, and Vít Baisa. 2024. Recognition of propaganda techniques in newspaper texts: Fusion of content and style analysis. *Expert Systems with Applications*, 251:124085.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. Mute: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd conference of the asia-pacific chapter of the association for computational linguistics and the 12th international joint conference on natural language processing: student research workshop*, pages 32–39.

Md Rajib Hossain, Sadia Afroze, Asif Ekbal, Mohammed Moshiul Hoque, and Nazmul Siddique. 2025. Multimodfusenet: Advancing multimodal text classification for low-resource languages through textual-visual feature fusion. *Knowledge-Based Systems*, 328:114085.

Chinmaya Hs and 1 others. 2021. Trollmeta@ dravidianlangtech-eacl2021: Meme classification using deep learning. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 277–280.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul NC, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 4948–4961.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, and 1 others. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Christos Koutlis, Manos Schinas, and Symeon Papadopoulos. 2023. Memetector: Enforcing deep focus for meme detection. *International Journal of Multimedia Information Retrieval*, 12(1):11.

Shiyi Li, Yike Wang, Liang Yang, Shaowu Zhang, and Hongfei Lin. 2024. Lmeme at semeval-2024 task 4: Teacher student fusion-integrating clip with llms for enhanced persuasion detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 628–633.

Mohamed Lichouri, Khaled Lounnas, Aicha Zitouni, Houda Latrache, and Rachida Djeradi. 2023. Usthb at araieval'23 shared task: Disinformation detection system based on linguistic feature concatenation. In *Proceedings of ArabicNLP 2023*, pages 508–512.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986.

Tarek Mahmoud and Preslav Nakov. 2024. Bertastic at semeval-2024 task 4: State-of-the-art multilingual propaganda detection in memes via zero-shot learning with vision-language models. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 503–510.

Monsoon NLP. 2024. Bangla-electra: A pretrained electra model for bengali. `https://huggingface.co/monsoon-nlp/bangla-electra`. Accessed: 2025-06-25.

Pir Ahmad Noman, Liu Yuanchao, Khursheed Aurangzeb, Muhammad Anwar Shahid, and Qazi Mazhar ul Haq. 2024. Semantic web-based propaganda text detection from social media using meta-learning. *Service Oriented Computing and Applications*.

Olumide E Ojo, Olaronke O Adebanji, Hiram Calvo, Damian O Dieke, Olumuyiwa E Ojo, Seye E Akinsanya, Tolulope O Abiola, and Anna Feldman. 2023. Legend at araieval shared task: Persuasion technique detection using a language-agnostic text representation model. *arXiv preprint arXiv:2310.09661*.

C.A. Piña-García. 2025. In-context learning for propaganda detection on twitter mexico using large language model meta ai. *Telematics and Informatics Reports*, 19:100232.

Jingnong Qu, Liunian Harold Li, Jieyu Zhao, Sunipa Dev, and Kai-Wei Chang. 2022. Disinfomeme: A multimodal dataset for detecting meme intentionally spreading out disinformation. *arXiv preprint arXiv:2205.12617*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Md. Tanvir Rahman, Shawly Ahsan, Jawad Hossain, Mohammed Moshiul Hoque, and M. Ali Akber Dewan. 2025. Multimodal emotion recognition system leveraging decision fusion with acoustic and visual cues. In *Pattern Recognition. ICPR 2024 International Workshops and Challenges*, pages 117–133, Cham. Springer Nature Switzerland.

Muhammad Umar Salman, Asif Hanif, Shady Shehata, and Preslav Nakov. 2023. Detecting propaganda techniques in code-switched social media text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16794–16812, Singapore. Association for Computational Linguistics.

Sagor Sarker. 2021. Banglabert: A bert-based language model for bengali. `https://huggingface.co/sagorsarker/bangla-bert-base`. Accessed: 2025-06-25.

Anik Mahmud Shanto, Mst. Sanjida Jamal Priya, Fahim Shakil Tamim, and Mohammed Moshiul Hoque. 2025. MDC^3: A novel multimodal dataset for commercial content classification in Bengali. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 311–320, Albuquerque, USA. Association for Computational Linguistics.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Zhivar Sourati, Vishnu Priya Prasanna Venkatesh, Darshan Deshpande, Himanshu Rawlani, Filip Ilievski, Hông Ân Sandlin, and Alain Mermoud. 2023. Robust and explainable identification of logical fallacies in natural language arguments. *Knowledge-Based Systems*, 266:110418.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR.

Ming-Hung Wang and Yu-Lin Chen. 2025. Beyond text: Detecting image propaganda on online social networks. *IEEE Transactions on Sustainable Computing*, 10(1):120–131.

Mohamed Zaytoon, Nagwa M El-Makky, and Marwan Torki. 2024. Alexunlp-mz at araieval shared task: contrastive learning, llm features extraction and multi-objective optimization for arabic multi-modal meme propaganda detection. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 512–517.

Yang Zhong and Bhiman Kumar Baghel. 2024. Multimodal understanding of memes with fair explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2007–2017.

A Annotation Guidelines

The following questions, tied to specific propaganda characteristics, guide the classification process, where memes that fall under one or more of these characteristics are considered propagandistic, and those that do not are treated as non-propagandistic:

- **Intent to Influence or Manipulate**: (i) Does the meme push a specific political, ideological, or social agenda? (ii) Does it encourage the viewer to act (support a cause, oppose a group, or adopt a belief)?
- **Emotional Appeal**: (i) Does the meme evoke strong emotions like fear, anger, pride, or sympathy? (ii) Does it use fear, exaggeration, or threats to influence opinions or actions?
- **Simplification of Complex Issues**: Does the meme reduce a complex issue to overly simple terms?
- **Polarization and Division**: Does the meme create an "us versus them" narrative?
- **Repetition and Catchphrases**: Does the meme repeat messages or use catchy slogans that stick in the audience's mind?
- **Misleading Information**: Does the meme include misinformation or disinformation?
- **Smear Tactics**: Does the content use negative claims to undermine the reputation of an individual or group without providing credible evidence?
- **Visual Symbolism and Transfer**: Does the meme use images or symbols (like national flags, religious icons, or culturally significant visuals) to evoke specific associations or emotions?

B Cohen's Kappa Score

The annotation quality for the MemeGuard dataset was assessed using Cohen's kappa coefficient among three undergraduate annotators, as shown in Table 8. Pairwise kappa scores were 0.85 (Annotator 1 & 2), 0.90 (Annotator 1 & 3), and 0.77 (Annotator 2 & 3), indicating significant to almost perfect agreement. The average kappa score of 0.84 shows the high consistency and reliability of the annotations, highlighting the robustness of the dataset for detecting Bengali propagandistic memes.

Pair	Kappa Score
P-1	0.85
P-2	0.90
P-3	0.77
Average	0.84

Table 8: Pairwise Cohen's kappa score

C Hyperparameters

Table 9 provides a detailed overview of the hyperparameters chosen for training both unimodal and multimodal baselines. All models were trained using these values, which we examined across every model.

Hyperparameter	Search Space
Batch Size	4, 8, 16
Epochs	10, 15, 20
Optimizer	Adam
Weight Decay	0.01, 0.1
Learning Rate	5e-4, (1, 2, 5) e-5

Table 9: Hyperparameters for all models