BanHateMe : Understanding Hate in Bangla Memes through Detection, Categorization, and Target Profiling

Md Ayon Mia[1], **Md Fahim**[2,3]

[1]*Dhaka International University*
[2]*Center for Computational & Data Sciences* [3]*Penta Global Limited*
Correspondence: {mdayonrahman100, fahimcse381}@gmail.com

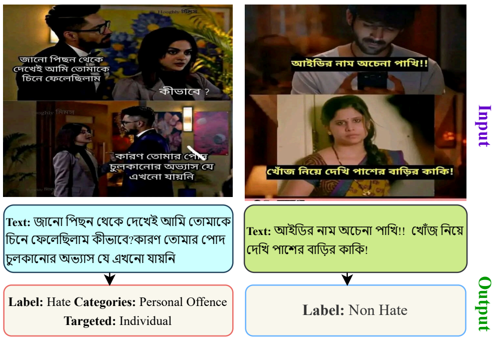
Abstract

Detecting hateful memes is a complex task due to the interplay of text and visuals, with subtle cultural cues often determining whether content is harmful. This challenge is amplified in Bangla, a low-resource language where existing resources provide only binary labels or single dimensions of hate. To bridge this gap, we introduce BanHateMe , a comprehensive Bangla hateful meme dataset with hierarchical annotations across three levels: binary hate, hate categories, and targeted groups. The dataset comprises 3,819 culturally grounded memes, annotated with substantial inter-annotator agreement. We further propose a hierarchical loss function that balances predictions across levels, preventing bias toward binary detection at the expense of fine-grained classification. To assess performance, we pair pretrained language and vision models and systematically evaluate three multimodal fusion strategies: summation, concatenation, and co-attention, demonstrating the effectiveness of hierarchical learning and cross-modal alignment. Our work establishes BanHateMe as a foundational resource for fine-grained multimodal hate detection in Bangla and contributes key insights for content moderation in low-resource settings. We release the code and dataset publicly at https://github.com/Ayon128/BanHateMe.

Disclaimer: This paper includes examples that may be offensive. Such content is presented only for research purposes and is unavoidable given the nature of the study.

1 Introduction

Memes have rapidly become one of the most influential forms of online communication, combining images with short text to convey ideas, humor, and social commentary. While often entertaining, they are also frequently used to spread hate in subtle and multimodal ways, embedding socio-political cues, cultural references, or stereotypes that can harm individuals and groups based on gender, politics, or religion. Their multimodal nature and concealed semantics make them especially difficult to analyze, as harmful meaning often arises from the interaction between visual and textual elements (Zannettou et al., 2018; Kiela et al., 2020). Moreover, meme content evolves dynamically with changing events, metaphors, and linguistic patterns, complicating detection even further (Pramanick et al., 2021). These challenges are particularly acute in low-resource languages such as Bangla, where multimodal hate has received little attention despite the language's widespread use online.



Text: জানো পিছন থেকে দেখেই আমি তোমাকে চিনে ফেলেছিলাম কীভাবে?কারণ তোমার পোদ চুলকানোর অভ্যাস যে এখনো যায়নি

Label: Hate **Categories:** Personal Offence **Targeted:** Individual

Text: আইডির নাম অচেনা পাখি!! খোঁজ নিয়ে দেখি পাশের বাড়ির কাকি!

Label: Non Hate

Figure 1: Given a meme image with associated text as input, the output is its hierarchical annotation: hateful memes are labeled with a hate category and targeted group (left), while non-hateful memes are marked as non hate (right).

To effectively assess hateful memes, it is not enough to simply determine whether the content is hateful. A comprehensive understanding requires capturing both the severity of hate through categories such as abusive, political, gender, personal offence, or religious, and the intended targets, including individuals, communities, organizations, or society. Existing Bangla meme datasets fall short in this regard: MUTE (Hossain et al., 2022)

Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 180–195
December 23, 2025 ©2025 Association for Computational Linguistics

Dataset	Task Type	Hate Cat	Target Groups	#Samples
Hossain et al., 2024b	Hate/Non-Hate & Target Entity Detection	✗	✓	7,148
Ahsan et al., 2024	Aggression Detection	✓	✗	4,848
Hossain et al., 2022	Hate vs. Non-Hate	✗	✗	4,158
Das and Mukherjee, 2023	Abusive vs. Non-Abusive	✗	✗	4,043
Ours	Hierarchical Classification	✓	✓	3,819

Table 1: Comparison of Bengali multimodal meme datasets. BANHATEME introduces hierarchical annotations, including binary labels, hate categories, and target groups, which are not jointly supported in prior resources.

provides only binary labels, BanglaAbuseMeme (Das and Mukherjee, 2023) focuses solely on abusive content, while more recent efforts such as BHM (Hossain et al., 2024b) and MIMOSA (Ahsan et al., 2024) annotate either categories or targets but never both. As summarized in Table 1, no current resource jointly supports all levels of analysis. To fill this gap, we introduce BANHATEME , a comprehensive Bangla multimodal hateful meme dataset with hierarchical annotation, where memes are classified as hateful or not, and hateful memes are further categorized by type and targeted group, thereby bridging this gap, as illustrated in Figure 1. We complement this design with a hierarchical loss function that balances predictions across levels, ensuring performance is not skewed toward binary detection at the expense of fine-grained recognition. Since Bangla lacks dedicated vision–language models, we combine pretrained language and vision encoders and systematically evaluate three fusion strategies—summation, concatenation, and co-attention to address alignment challenges. Our experiments show that BanglaBERT with Swin Transformer and concatenation yields competitive results, while co-attention provides clear improvements in target group recognition. These findings underscore the importance of hierarchical modeling and cross-modal alignment for multimodal hate detection in Bangla. Our key contributions are as follows:

- We introduce BANHATEME , the first Bangla hateful meme dataset with hierarchical annotations across binary labels, hate categories, and targeted groups.
- Propose a hierarchical loss function to balance supervision across multiple levels.
- Conduct a comprehensive evaluation comparing three multimodal fusion strategies, demonstrating the effectiveness of hierarchical learning for Bangla multimodal hate detection.

2 Related Work

2.1 Hateful memes dataset.

The release of the Hateful Memes Challenge (Kiela et al., 2020) established a benchmark for multimodal hate detection, highlighting the need for joint reasoning across text and images. Since then, several English datasets have expanded the space, including large-scale resources for offensive or harmful memes (Suryawanshi et al., 2020; Gomez et al., 2020; Pramanick et al., 2021). Efforts have also extended to low-resource languages, such as Hindi (Kumari et al., 2023; Rajput et al., 2022) and Greek (Perifanos and Goutsos, 2021). For Bangla, multimodal resources remain limited. MUTE (Hossain et al., 2022) introduced 4,158 memes labeled for binary hate detection, while BanglaAbuseMeme (Das and Mukherjee, 2023) provided 4,043 abusive memes. More recently, BHM (Hossain et al., 2024b) added target entity annotations (e.g., Individual, Organization, Community, Society), and MIMOSA (Ahsan et al., 2024) focused on aggression-specific categories such as Political, Gender, and Religious. BANMIME (Mia et al., 2025) further contributed to the Bangla multimodal landscape by introducing 2,000 misogynistic memes annotated with metaphor localization and human-written explanations. Beyond these resources, ExMUTE (Debnath et al., 2025) expands Bangla multimodal hate research by incorporating contextual labels across religion, politics, gender, and other domains, demonstrating the importance of context-aware annotations for improving hateful meme understanding.

2.2 Hateful memes detection methods.

Research on multimodal hateful memes has explored a range of fusion techniques. Conventional fusion approaches concatenate text and image features to form a joint representation (Vijayaraghavan et al., 2021; Gomez et al., 2020). Some works adopted bilinear pooling (Chandra et al., 2021),

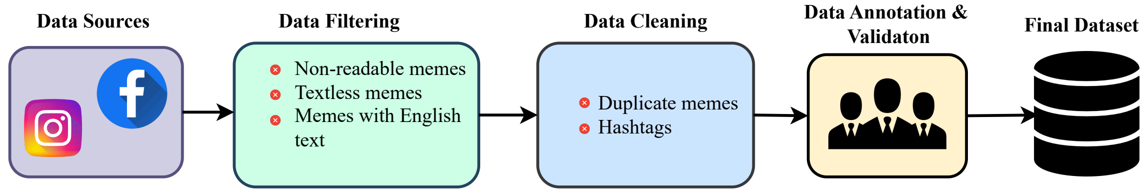


Figure 2: BANHATEME dataset development pipeline showing data collection from social media platforms, filtering to discard irrelevant content, cleaning to remove duplicates and extraneous elements, and annotation with validation to construct the final dataset.

while others fine-tuned vision–language transformer architectures such as ViLBERT, MMBT, and VisualBERT (Kiela et al., 2020). More recent studies explored prompting techniques for English hateful memes (Cao et al., 2023). Despite these advances, aligning textual and visual features remains underexplored, even though effective feature alignment is critical for robust multimodal representations (Zeng et al., 2021; Liu et al., 2019). In the Bangla context, emerging approaches such as Align-before-Attend (Hossain et al., 2024a) and Multimodal Attentive Fusion (Ahsan et al., 2024) demonstrate the value of improved cross-modal integration for meme classification.

3 BANHATEME : Dataset Creation

Following the limitations of existing Bangla hate meme datasets highlighted in Table 1, our dataset design focuses on enriching memes with layered annotations that move beyond binary hate labels (Hate vs. Non-Hate) to include five hate categories and four targeted group types. Particular attention was given to sourcing content from diverse platforms, capturing cultural and social nuances during annotation. The overall dataset construction pipeline is illustrated in Figure 2.

Data Collection. We collected memes from publicly accessible Bangla-speaking communities on major social media platforms, primarily Facebook and Instagram, between April 2022 and May 2025. We used search terms such as "Bangla memes", "Bangla hate memes", "Bangla abusive memes", "Bangla political memes", etc. to identify relevant meme sources. To comply with copyright and ethical standards, only memes from open groups, public pages, and non-private sources were included. In total, 5,560 memes were initially collected, with Facebook contributing 3,562 samples and Instagram providing the remaining 1,998.

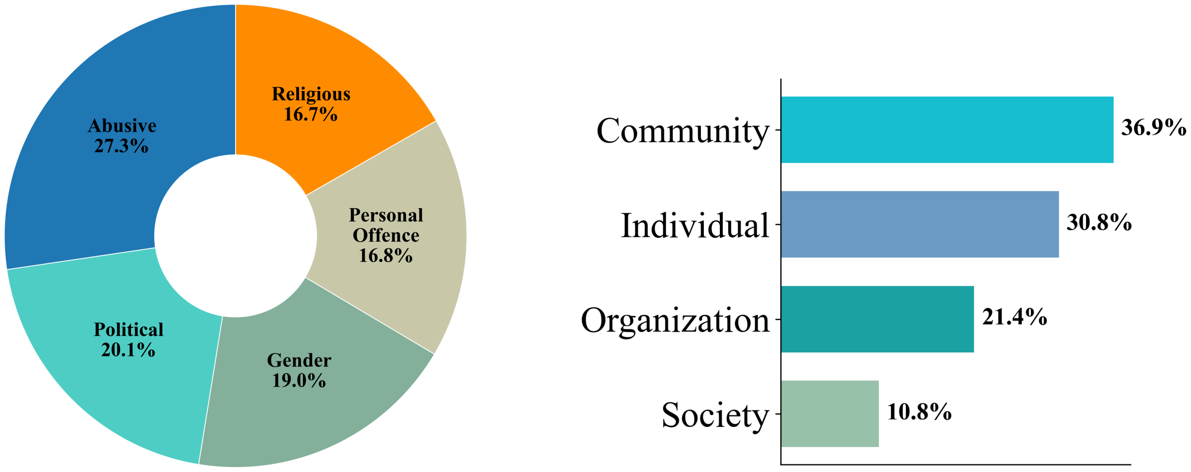
Data Filtering. To ensure the quality and relevance of our dataset, we applied a rigorous filtering procedure. Specifically, we removed: (1) memes with unreadable visual or textual content, (2) memes lacking any textual information, and (3) memes containing only English text, as our study emphasizes Bangla linguistic and cultural markers. This process resulted in the removal of 1,012 memes, leaving the 4,548 samples with extractable, readable Bangla text for subsequent analysis.

Data Cleaning. During this stage, we removed duplicate memes that appeared across different sources. We also stripped away non-informative textual elements, such as hashtags, which could add noise without contributing to the semantic analysis. After this cleaning procedure, 729 redundant and extraneous entries were discarded, leaving a final dataset of 3,819 text-bearing memes ready for annotation.

Text Extraction. Existing OCR systems perform poorly on Bangla text embedded in images, often producing noisy or incomplete outputs. To ensure accuracy, we relied on manual transcription of meme text. Two native typists were recruited to carry out the task, with the dataset evenly divided between them. Each typist was compensated at a rate of 1.5 BDT per sample.

Data Annotation. To annotate the BANHATEME dataset, we hired three Bangla-speaking undergraduate annotators with strong familiarity with local meme culture and online discourse. Their background in meme creation and cultural interpretation enabled them to recognize subtle hateful cues that might otherwise be overlooked. The annotators were provided with detailed annotation guidelines (Appendix Section A) and were instructed to complete the task within 25 days.

Annotation followed a two-stage hierarchical process. In the first stage, annotators determined whether each meme should be labeled as Hate or Not Hate. In the second stage, hateful memes were further classified into five hate categories—Abusive, Political, Gender, Personal Of-



(a) Hate Categories

(b) Targeted Groups

Figure 3: Distribution of hateful memes in the BanHateMe dataset across (a) hate categories and (b) targeted groups.

	Label	Kappa(κ)	Avg.
Primary	Hate	0.81	0.78
	Non Hate	0.75	
Hate Categories	Abusive	0.64	0.68
	Political	0.70	
	Personal	0.66	
	Gender	0.71	
	Religious	0.69	
Targeted Groups	Individual	0.71	0.69
	Organization	0.69	
	Community	0.68	
	Society	0.67	

Table 2: Inter-annotator agreement for the BanHateMe dataset, measured using Cohen's kappa (κ) across the binary task, hate categories, and targeted groups, with averages indicating substantial reliability.

fence, and Religious, followed by the study (Haider et al., 2024). Each hateful meme was also tagged with one of four targeted group types—Community, Individual, Organization, or Society, inspired by the work (Hossain et al., 2024b). All memes were annotated independently by three annotators, and final labels are assigned through majority voting to reduce individual bias. Annotators were compensated at a rate of 2 BDT per sample.

Data Validation. We assessed inter-annotator reliability using Cohen's kappa (κ) for each level of the hierarchical labeling task, with detailed results presented in Table 2. Overall, the scores indicate substantial agreement for the binary classification as well as across both hate categories and targeted groups. No score is less than 0.64 which indicates a good agreement between the annotators. The BanHateMe dataset serves as a trustworthy resource for Bangla memes.

Source Distribution	# Samples
Facebook	2517
Instagram	1302
Splits	
- Train	2673
- Val	381
- Test	765
Text Statistics	
Max Character Length	611
Mean Character Length	82.36
Min Character Length	10
Max Word Count	111
Mean Word Count	14.35
Min Word Count	3

Table 3: Statistical overview of the BanHateMe dataset, showing source distribution, data splits, and text statistics.

4 BanHateMe : Dataset Statistics

Meme Collection. The BanHateMe dataset consists of 3,819 labeled Bangla memes collected from two major platforms: Facebook (2,517) and Instagram (1,302). To enable systematic evaluation, we applied a stratified 70–10–20 split, resulting in training (2,673), validation (381), and test (765) partitions while preserving the overall distribution of labels. The dataset exhibits considerable linguistic diversity, with meme texts ranging from 10 to 611 characters and 3 to 111 words, averaging 82.36 characters and 14.35 words per instance. Table 3 summarizes the dataset statistics.

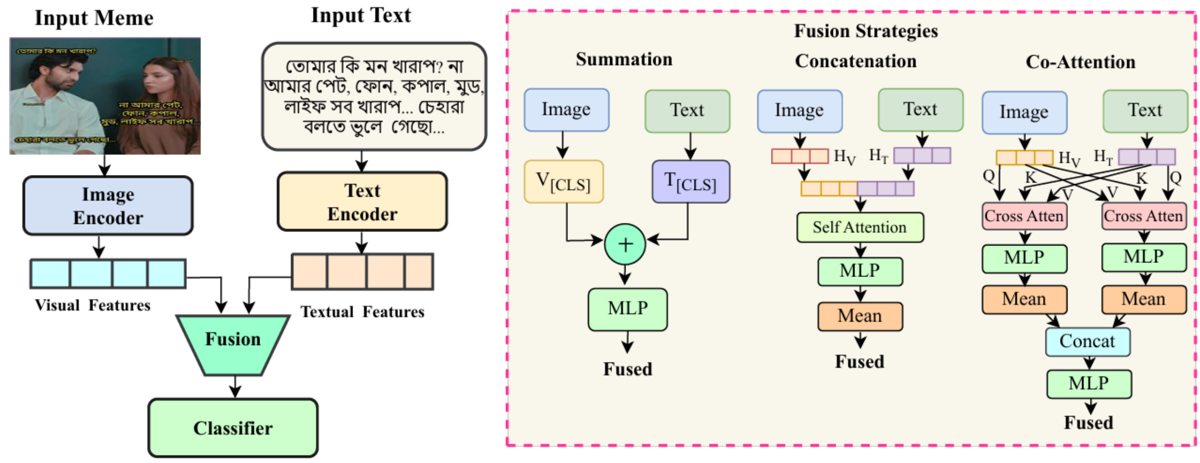


Figure 4: Overview of our hierarchical multimodal framework. Image and text are encoded separately, and their representations are fused using summation, concatenation, or co-attention. The fused features are then used for hierarchical classification across binary labels, hate categories, and targeted groups.

Label Distribution. The BANHATEME dataset contains 2,050 Non-Hate memes and 1,769 Hate memes. Among the hateful memes, Figure 3(a) presents the distribution across hate categories, where Abusive content is most frequent, followed by Political and Gender-based memes, while Religious and Personal Offence are comparatively less common. Figure 3(b) illustrates the breakdown of targeted groups, showing that Communities are the most frequent targets, followed by Individuals and Organizations, with Society-level hate being the least represented.

5 Methodology

Our approach leverages the multimodal nature of memes, which combine visual and textual information to convey meaning. Each meme in our dataset is treated as a multimodal input $x = (x_V, x_T)$, where x_V represents the image content and x_T denotes the extracted text from the meme.

As illustrated in Figure 4, we process each modality through dedicated encoders to obtain modality-specific representations. These representations are then fused via a fusion module to produce a combined multimodal embedding, which is passed into a classification module for prediction. Our implementation utilizes a hierarchical classification loss to improve performance.

5.1 Modality-Specific Representation

To extract meaningful features, we use pretrained encoders tailored to each modality. The image input x_V is processed by a transformer-based vision encoder ϕ_V, which splits the image into patches and appends a special `[CLS]` token representing the entire image. The output is:

$$H_V = \{v_{[\text{CLS}]}, v_1, \ldots, v_m\} = \phi_V(x_V)$$

Similarly, the extracted text x_T is fed into a pretrained text encoder ϕ_T, producing token embeddings including the `[CLS]` token:

$$H_T = \{t_{[\text{CLS}]}, t_1, \ldots, t_n\} = \phi_T(x_T)$$

5.2 Modality Fusion

The representations from both modalities are combined using one of three fusion strategies:

Summation-Based Fusion We sum the `[CLS]` embeddings from both modalities and pass the result through an MLP:

$$h_{\text{fused}} = \text{MLP}(v_{[\text{CLS}]} + t_{[\text{CLS}]})$$

Concatenation-Based Fusion The token embeddings from both modalities are concatenated and processed by a self-attention block followed by an MLP. The fused sequence is mean-pooled to obtain a fixed-length vector:

$$\begin{aligned} H_{\text{fused}} &= \text{MLP}\big(\text{Self-Attention}([H_V; H_T])\big) \\ h_{\text{fused}} &= \text{Mean-Pooling}(H_{\text{fused}}) \end{aligned}$$

Co-Attention Based Fusion Co-attention computes inter-modal attention by using queries from one modality and keys/values from the other. This yields two fused outputs:

Fusion Method	Non-Hate				Hate				Overall	
	P	R	F1	Acc	P	R	F1	Acc	F1	Acc
BanglaBERT + ViT										
Sum based	69.58	81.27	74.97	81.27	72.98	58.76	65.10	58.76	70.40	70.85
Concatenation	69.94	79.81	74.55	79.81	71.96	60.17	65.54	60.17	70.38	70.72
Co-Attention	69.27	66.91	68.07	66.91	63.04	65.54	64.27	65.54	66.31	66.27
BanglaBERT + Swin										
Sum based	72.25	82.96	77.24	82.97	76.11	62.99	68.93	62.99	73.39	73.73
Concatenation	72.31	**85.15**	**78.21**	**85.16**	**78.29**	62.15	**69.29**	62.15	**74.08**	**74.51**
Co-Attention	69.48	74.21	71.76	74.21	67.48	62.15	64.71	62.15	68.50	68.63
XLM-RoBERTa + ViT										
Sum based	62.03	51.68	56.39	51.69	63.62	**72.75**	67.88	**72.75**	62.56	63.01
Concatenation	65.04	64.71	64.88	64.72	59.27	59.60	59.44	59.60	62.36	62.35
Co-Attention	65.71	61.06	63.30	61.07	58.22	62.99	60.52	62.99	62.01	61.96
XLM-RoBERTa + Swin										
Sum based	69.12	82.22	75.11	82.24	73.55	57.33	64.44	57.34	70.18	70.72
Concatenation	**72.75**	76.63	74.64	76.64	71.08	66.66	68.80	66.67	71.94	72.03
Co-Attention	65.05	81.50	72.35	81.51	69.60	49.14	57.62	49.15	65.53	66.54

Table 4: Model benchmarking results on the test split of the BanHateMe dataset are reported. Here, P, R, F1, and Acc represent Precision, Recall, F1 Score, and Accuracy, respectively.

$$H_V^{\text{fused}} = \text{MLP}\left(\sigma\left(\frac{(W_Q H_V)(W_K H_T)^\top}{\sqrt{d_k}}\right)(W_V H_T)\right)$$

$$H_T^{\text{fused}} = \text{MLP}\left(\sigma\left(\frac{(W_Q H_T)(W_K H_V)^\top}{\sqrt{d_k}}\right)(W_V H_V)\right)$$

where W_Q, W_K, W_V are learned projection matrices, σ is the softmax function, and d_k is a scaling factor.

Mean-pooling is applied to each fused representation, which are then concatenated and passed through an MLP to get the final fused vector:

$$\begin{aligned} h_V^{\text{fused}} &= \text{Mean-Pooling}(H_V^{\text{fused}}) \\ h_T^{\text{fused}} &= \text{Mean-Pooling}(H_T^{\text{fused}}) \\ h_{\text{fused}} &= \text{MLP}\left([h_V^{\text{fused}}; h_T^{\text{fused}}]\right) \end{aligned}$$

This approach captures detailed interactions between the image and text modalities.

5.3 Classification Module

The fused representation h_{fused}, is passed into a classification head to generate the prediction logits. This classifier is implemented as a simple linear transformation, where the logits are computed using the following operation:

$$\text{logits} = W_{\text{C}} \cdot h_{\text{fused}}$$

5.4 Hierarchical Loss

To train the model in accordance with the hierarchical structure of the classification task, we design a composite loss function. At the core of this formulation is a binary cross-entropy loss, $\mathcal{L}_{\text{binary}}$, which measures the performance of the model in distinguishing between hateful and non-hateful memes.

Once a meme is predicted as hateful, two additional cross-entropy losses are computed. The first, denoted as $\mathcal{L}_{\text{hate\_cat}}$, corresponds to the hate category classification. The second, $\mathcal{L}_{\text{target\_grp}}$, evaluates the model's ability to correctly identify the target group affected by the hate content. The total loss function combines these three components in a weighted manner:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{binary}} + \alpha \cdot \mathcal{L}_{\text{hate\_cat}} + \beta \cdot \mathcal{L}_{\text{target\_grp}}$$

In this equation, α and β are hyperparameters that control the contribution of the hate category and target group classification losses, respectively.

5.5 Experiment Setup

We conducted all experiments on the Kaggle platform using an NVIDIA Tesla P100 GPU with 16 GB VRAM, 32 GB RAM, and 8 CPU cores. For text encoding, we employed two pretrained language models, BanglaBERT (Bhattacharjee et al., 2021) and RoBERTa (Conneau et al., 2019), while for image encoding we used ViT (Dosovitskiy et al., 2020) and Swin Transformer (Liu et al., 2021). We selected these text and vision encoders as they have demonstrated strong performance in Bangla NLP

Fusion Method	Hate Category						Target Group				
	Ab	Po	Ge	Per	Re	Avg	Co	Ind	Org	So	Avg
BanglaBERT + ViT											
Sum based	60.87	64.29	31.03	25.40	32.88	42.89	44.86	58.23	59.20	09.23	42.88
Concatenation	65.74	71.53	52.57	15.79	61.54	53.43	61.32	57.76	63.24	08.13	47.61
Co-Attention	67.23	64.75	40.30	24.24	33.85	46.07	66.67	62.01	67.63	**48.78**	61.27
BanglaBERT + Swin											
Sum based	54.02	**77.61**	32.99	31.84	**76.47**	54.59	59.11	58.25	**78.57**	7.14	50.77
Concatenation	**69.53**	75.56	53.89	19.18	74.00	**58.43**	69.63	56.45	74.45	12.32	53.21
Co-Attention	66.67	71.76	56.65	25.87	66.67	57.52	**71.27**	**65.71**	71.11	38.64	61.68
XLM-RoBERTa + ViT											
Sum based	54.27	45.40	43.98	37.62	10.34	38.32	64.67	60.50	07.23	6.22	34.66
Concatenation	65.38	50.00	45.45	**38.71**	25.35	44.98	63.26	63.76	25.00	29.03	45.26
Co-Attention	64.73	45.83	38.34	18.18	34.48	40.31	61.62	57.00	08.92	32.84	40.10
XLM-RoBERTa + Swin											
Sum based	62.46	70.83	50.37	08.21	68.09	51.99	64.29	60.39	70.27	31.33	56.57
Concatenation	68.80	75.36	50.00	22.50	72.00	57.73	66.94	62.01	77.03	45.78	**62.94**
Co-Attention	63.26	70.34	**57.83**	26.51	64.65	56.52	70.80	65.09	71.14	38.36	61.35

Table 5: Performance across hate categories and target groups on the test split of the BANHATEME dataset, reported using F1 score. Here, Ab, Po, Ge, Per, and Re refer to Abusive, Political, Gender, Personal Offence, and Religious categories, while Co, Ind, Org, and So denote Community, Individual, Organization, and Society, respectively.

and multimodal classification tasks. The MLP layers operated on a 768-dimensional representation, and cross-attention modules also yielded aligned multimodal features of 768 dimensions. All models were fine-tuned for up to 10 epochs with early stopping to prevent overfitting, using a batch size of *16* and learning rate of *2e-5*. For hierarchical loss, we applied weighting parameters with α = *0.2, 0.8, 0.5, 1.0* and β = *0.8, 0.2, 0.5,1.0* across different levels. Our implementation relied on HuggingFace *Transformers 4.45.1* (Wolf et al., 2020) with *PyTorch 2.4.0* as the backend, and we used *NumPy 1.26.4*, *Pandas 2.2.3*, *Matplotlib 3.7.5*, *Seaborn 0.12.2*, and *scikit-learn 1.2.2* for data processing, analysis, and visualization.

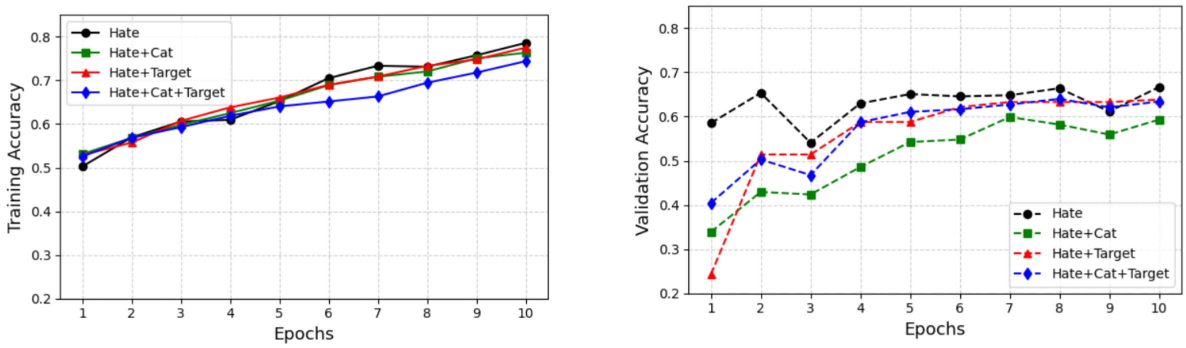
6 Results and Analysis

The performance of different configurations of language and vision models on hate/non-hate detection, hate categories, and targeted groups is reported in Tables 4 and 5. We analyze the results along the following dimensions:

Impact of Language Model. BanglaBERT consistently outperformed XLM-RoBERTa in overall binary detection, achieving gains of about 3-5% in both F1 and accuracy. For hate categories, BanglaBERT showed clear improvements in Political, Religious, and Abusive memes, where performance increased by roughly 4-6%. In contrast, XLM-RoBERTa performed better in Personal Offence and Gender. For targeted groups, BanglaBERT delivered stronger results in Community and Individual prediction, while XLM-RoBERTa achieved higher scores for Organization. Overall, the outcomes indicate that monolingual models are more effective at capturing broad linguistic and cultural cues in Bangla.

Impact of Vision Model. The vision backbone played a crucial role in shaping overall performance. Swin consistently outperformed ViT across most configurations, yielding relative gains of about 2–3% in binary hate detection and up to 6–8% in categories such as Political and Religious. For targeted groups, Swin provided clear advantages in detecting Individuals and Organizations, with improvements ranging from 7–10% over ViT. An exception emerged in Personal Offence, where ViT combined with XLM-RoBERTa achieved a notable score across all settings. Nonetheless, Swin remained the more reliable encoder overall, highlighting the importance of localized visual representations for domains where subtle contextual markers drive hateful interpretation.

Impact of Fusion Strategy. Fusion strategies influenced the performance of the models. For binary detection, concatenation proved most effective, giving BanglaBERT with Swin the highest overall scores and surpassing summation and co-attention by 1–4%. For hate categories, the best method varied: concatenation excelled in Abusive and Personal, summation performed better in Political and Religious, while co-attention achieved the top re-



(a) Training accuracy across categorical cross-entropy loss variants

(b) Validation accuracy across categorical cross-entropy loss variants

Figure 5: Impact of hierarchical loss on model performance, showing smoother convergence and reduced variance compared to single-task and partial supervision settings.

Value of α & β	$\alpha = 0.2, \beta = 0.8$			$\alpha = 0.5, \beta = 0.5$			$\alpha = 0.8, \beta = 0.2$			$\alpha = 1, \beta = 1$		
	H/NH	**Cat**	**Tar**	**H/NH**	**Cat**	**Tar**	**H/NH**	**Cat**	**Tar**	**H/NH**	**Cat**	**Tar**
BanglaBERT + ViT												
Sum based	72.05	51.89	58.75	70.40	42.89	42.88	69.96	56.39	58.75	68.25	49.33	48.06
Concatenation	70.69	53.25	47.84	70.38	53.43	47.61	70.38	53.43	45.58	69.57	59.93	63.42
Co-Attention	67.02	56.39	58.75	66.31	46.07	61.27	67.44	58.68	63.21	65.71	57.01	60.55
BanglaBERT + Swin												
Sum based	73.39	54.59	50.77	73.39	54.59	50.77	70.63	54.26	48.34	66.44	60.57	53.42
Concatenation	70.91	57.10	49.64	74.51	58.43	53.21	72.03	60.40	59.24	66.89	36.77	42.49
Co-Attention	69.78	55.78	61.89	68.50	57.50	61.68	63.97	50.87	49.09	66.99	55.69	63.72
XLM-RoBERTa + ViT												
Sum based	62.73	36.72	31.44	62.56	51.99	56.57	63.53	40.35	31.20	65.65	61.63	65.64
Concatenation	63.84	40.13	37.78	62.36	44.98	45.26	61.29	46.45	36.06	64.16	57.33	54.42
Co-Attention	62.20	36.06	40.48	62.01	40.31	40.10	62.68	36.55	37.42	66.06	49.88	58.11
XLM-RoBERTa + Swin												
Sum based	61.99	52.16	51.59	70.18	51.99	56.57	62.87	53.87	49.87	62.78	23.02	33.23
Concatenation	70.37	57.13	51.54	71.94	57.73	62.94	72.77	63.58	56.04	66.24	57.08	51.74
Co-Attention	63.79	51.02	58.97	65.53	56.52	61.35	64.05	50.23	56.10	64.63	52.93	58.47

Table 6: Impact of α and β in the hierarchical loss. We report the overall performance (F1 score) of detecting hate/non-hate, along with hate category and target group prediction. Here H/NH, Cat, and Tar refer to Hate/Non-Hate, Hate Category, and Target Group prediction results, respectively.

sult in Gender. For targeted groups, co-attention consistently led for Community, Individual, and Society, whereas summation was best for Organization.

Impact of Hierarchical Loss. Figures 5(a) and 5(b) compare models trained on the binary Hate vs. Non-Hate task under different loss formulations. Using only categorical cross-entropy on the binary task yields the highest training accuracy but quickly saturates on validation. Incorporating auxiliary supervision from categories (Hate+Cat) or targets (Hate+Target) slightly reduces training accuracy but stabilizes validation curves. The full hierarchical loss (Hate+Cat+Target) achieves the most consistent validation accuracy with reduced variance across epochs.

Impact of α and β. As shown in Table 6, varying the values of α and β in the hierarchical loss influences the trade-off between hate category and target group prediction. While higher weight on one component improves that sub-task, it typically reduces the other. We find that setting $\alpha = 0.5$ and $\beta = 0.5$ provides the most effective overall performance in terms of F1 score.

Error Analysis. We conduct both quantitative and qualitative error analyses to better understand model behavior across binary, category, and target group levels. A detailed analysis is provided in Appendix C.

7 Conclusion

We present BanHateMe , Bangla hateful meme dataset with hierarchical annotations covering binary labels, hate categories, and targeted groups.

This resource advances multimodal hate detection in low-resource settings by providing culturally grounded annotations and a hierarchical loss that balances predictions across levels with different fusion techniques and their impact on the modality alignment. We envision BANHATEME as a foundation for building culturally aware multimodal moderation systems in Bangla and as a catalyst for future research on hierarchical modeling, fusion strategies, and cross-modal reasoning in underrepresented languages.

Limitations

A primary limitation of our study is the relatively modest dataset size, constrained by the effort required for high-quality hierarchical annotations across categories and targeted groups. While this scale provides a strong foundation, larger datasets would be necessary to further improve generalizability. Another limitation lies in the reliance on pretrained language and vision encoders not originally optimized for Bangla multimodal content. As a result, current models struggle with subtle cultural cues and fine-grained cross-modal interactions. In future work, we plan to expand the dataset through more efficient annotation strategies and explore culturally adapted multimodal architectures tailored to Bangla content. Previous studies on Bangla and multilingual languages (Haider et al., 2024; Fahim et al., 2024; Ahmed et al., 2024) have observed performance variations in large language models (LLMs) across different prompting techniques. In future, we also plan to experiment LVLMs using different techniques to see the impact of the prompt on the performance of LVLMs in our dataset.

Ethical Statement

We collected memes exclusively from publicly accessible social media sources and excluded any content containing explicit nudity or personally identifiable information (PII). All memes were manually reviewed to remove duplicates, unreadable content, or irrelevant material. Annotators were Bangla-speaking individuals familiar with online discourse, and their privacy was strictly maintained; no personal data about them was collected or shared. They were fairly compensated for their work at rates consistent with local norms. To mitigate potential biases, we developed comprehensive annotation guidelines, employed a multi-stage review process, and utilized majority voting to resolve disagreements. Nevertheless, we acknowledge that subjective judgments in hate classification may introduce residual biases. Our dataset is intended solely for research on multimodal hate detection in low-resource languages and should not be misused for malicious purposes. All resources will be released publicly to foster transparency, reproducibility, and future research. We emphasize that any harmful stereotypes or biases in the dataset are unintentional, and we have no intent to harm any individual or community.

References

Fahim Ahmed, Md Fahim, Md Ashraful Amin, Amin Ahsan Ali, and AKM Rahman. 2024. Improving the performance of transformer-based models over classical baselines in multiple transliterated languages. In *ECAI 2024*, pages 4043–4050. IOS Press.

Shawly Ahsan, Eftekhar Hossain, Omar Sharif, Avishek Das, Mohammed Moshiul Hoque, and M Dewan. 2024. A multimodal framework to detect target aware aggression in memes. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2487–2500.

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*.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2023. Prompting for multimodal hateful meme classification. *arXiv preprint arXiv:2302.04156*.

Mohit Chandra, Dheeraj Pailla, Himanshu Bhatia, Aadilmehdi Sanchawala, Manish Gupta, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021. “subverting the jewtocracy”: Online antisemitism detection using multimodal deep learning. In *Proceedings of the 13th ACM Web Science Conference 2021*, pages 148–157.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Mithun Das and Animesh Mukherjee. 2023. Banglaabusememe: A dataset for bengali abusive meme classification. *arXiv preprint arXiv:2310.11748*.

Riddhiman Swanan Debnath, Nahian Beente Firuj, Abdul Wadud Shakib, Sadia Sultana, and Md Saiful Islam. 2025. Exmute: A context-enriched multimodal dataset for hateful memes. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 83–89.

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

Md Fahim, Fariha Tanjim Shifat, Fabiha Haider, Deeparghya Dutta Barua, MD Sakib Ul Rahman Sourove, Md Farhan Ishmam, and Md Farhad Alam Bhuiyan. 2024. Banglatlit: A benchmark dataset for back-transliteration of romanized bangla. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14656–14672.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478.

Fabiha Haider, Fariha Tanjim Shifat, Md Farhan Ishmam, Deeparghya Dutta Barua, Md Sakib Ul Rahman Sourove, Md Fahim, and Md Farhad Alam. 2024. Banth: A multi-label hate speech detection dataset for transliterated bangla. *arXiv preprint arXiv:2410.13281*.

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.

Eftekhar Hossain, Omar Sharif, Mohammed Moshiul Hoque, and Sarah M Preum. 2024a. Align before attend: Aligning visual and textual features for multimodal hateful content detection. *arXiv preprint arXiv:2402.09738*.

Eftekhar Hossain, Omar Sharif, Mohammed Moshiul Hoque, and Sarah M Preum. 2024b. Deciphering hate: identifying hateful memes and their targets. *arXiv preprint arXiv:2403.10829*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Gitanjali Kumari, Dibyanayan Bandyopadhyay, and Asif Ekbal. 2023. Emoffmeme: identifying offensive memes by leveraging underlying emotions. *Multimedia Tools and Applications*, 82(29):45061–45096.

Fenglin Liu, Yuanxin Liu, Xuancheng Ren, Xiaodong He, and Xu Sun. 2019. Aligning visual regions and textual concepts for semantic-grounded image representations. *Advances in Neural Information Processing Systems*, 32.

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.

Md Ayon Mia, Akm Moshiur Rahman Mazumder, Khadiza Sultana Sayma, Md Fahim, Md Tahmid Hasan Fuad, Muhammad Ibrahim Khan, and Akmmahbubur Rahman. 2025. Banmime: Misogyny detection with metaphor explanation on bangla memes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17824–17850.

Konstantinos Perifanos and Dionysis Goutsos. 2021. Multimodal hate speech detection in greek social media. *Multimodal Technologies and Interaction*, 5(7):34.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Detecting harmful memes and their targets. *arXiv preprint arXiv:2110.00413*.

Kshitij Rajput, Raghav Kapoor, Kaushal Rai, and Preeti Kaur. 2022. Hate me not: detecting hate inducing memes in code switched languages. *arXiv preprint arXiv:2204.11356*.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 32–41.

Prashanth Vijayaraghavan, Hugo Larochelle, and Deb Roy. 2021. Interpretable multi-modal hate speech detection. *arXiv preprint arXiv:2103.01616*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the origins of memes by means of fringe web communities. In *Proceedings of the internet measurement conference 2018*, pages 188–202.

Yan Zeng, Xinsong Zhang, and Hang Li. 2021. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*.

A Annotation Guidelines

This section outlines the guidelines developed for annotating the BANHATEME dataset. Annotators examined both visual and textual elements of each meme to determine whether it should be labeled as Hate or Non-Hate. If annotated as Hate, the meme was further classified into one hate category (Abusive, Political, Gender, Personal offence, Religious) and one targeted group type (Community, Individual, Organization, Society). Figure 6 illustrates representative examples of Hate and Non-Hate memes, along with their assigned categories and target groups.

A.1 General Instructions

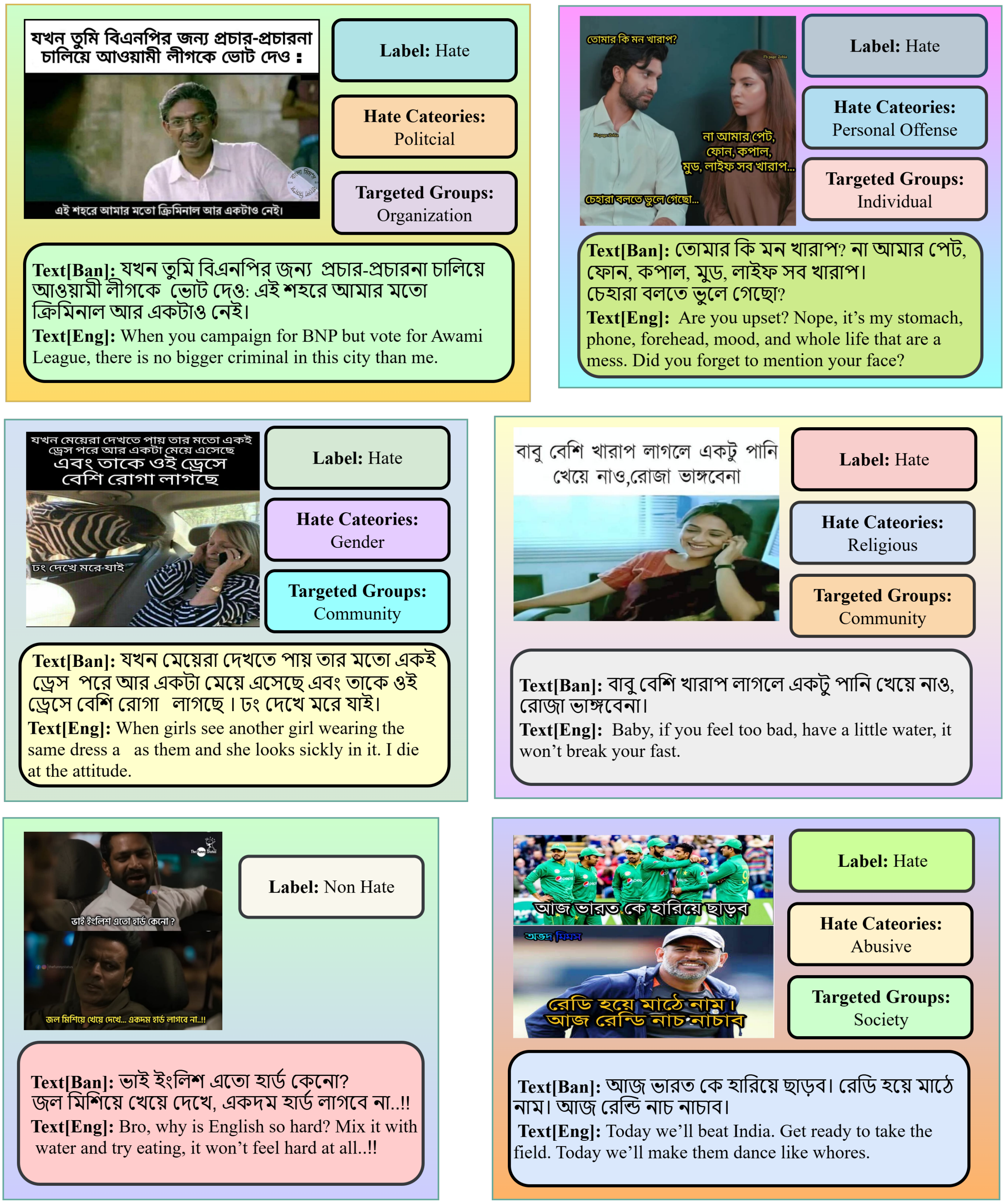
The following instructions guided the annotation process:

- **Multimodal Analysis:** Annotators jointly analyzed text and image, considering semantic interplay and cultural context.
- **Hierarchical Labeling:** Each meme was first classified as Hate or Non-Hate. Hate memes were further assigned exactly one hate category and one targeted group type.
- **Annotation Consistency:** Definitions were applied uniformly across the dataset to ensure reliability and reproducibility.
- **Cultural Relevance:** Bangla-specific expressions, code-mixed text, and culturally grounded memes were carefully interpreted.

A.2 Categories and Definitions

Our taxonomy captures five categories of hateful content commonly expressed in Bangla memes. Each category is defined below, along with representative indicators to assist annotation.

- **Abusive**
  Definition: Content that employs profane, offensive, or degrading language intended to insult, belittle, or provoke hostility, without making explicit threats of physical harm
  Indicators:
  - Use of curse words, slurs, or offensive profanity aimed at individuals or groups
  - Persistent use of hostile, degrading, or insulting language
  - Ridicule or mockery expressed in an aggressive manner, designed to provoke anger, humiliation, or distress
- **Political**
  Definition: Content that targets people or groups based on their political views or affiliations, often through incitement, hostility, or derogatory framing.
  Indicators:
  - Use of dehumanizing metaphors to describe political opponents
  - Explicit calls for violence or exclusion against political group
  - Spread of hostile disinformation to provoke hate or polarization
- **Gender**
  Definition: Content that demeans, stereotypes, or excludes individuals on the basis of gender or gender identity. This includes misogynistic, misandrist, or transphobic expressions that normalize discrimination or gender-based violence.
  Indicators:
  - Use of sexist or patriarchal stereotypes in jokes or insults
  - Justification, encouragement, or trivialization of gender-based violence
  - Derogatory remarks aimed at women, men, or gender-diverse identities
- **Personal Offence**
  Definition: Content that delivers targeted insults or demeaning remarks toward a specific individual, often exploiting personal vulnerabilities, traits, or experiences to humiliate or attack.
  Indicators:
  - Use of derogatory nicknames, personal slurs, or demeaning epithets
  - Mockery of an individual's tragedy, disability, or personal hardship
  - Attacks ridiculing someone's physical appearance or lifestyle
- **Religious**
  Definition: Content that marginalizes or demonizes individuals or communities on the


যখন তুমি বিএনপির জন্য প্রচার-প্রচারনা চালিয়ে আওয়ামী লীগকে ভোট দেও :
এই শহরে আমার মতো ক্রিমিনাল আর একটাও নেই।
তোমার কি মন খারাপ?
না আমার পেট, ফোন, কপাল, মুড, লাইফ সব খারাপ...
চেহারা বলতে ভুলে গেছো...
যখন মেয়েরা দেখতে পায় তার মতো একই ড্রেস পরে আর একটা মেয়ে এসেছে এবং তাকে ওই ড্রেসে বেশি রোগা লাগছে
ঢং দেখে মরে যাই
বাবু বেশি খারাপ লাগলে একটু পানি খেয়ে নাও,রোজা ভাঙ্গবেনা
ভাই ইংলিশ এতো হার্ড কেনো ?
জল মিশিয়ে খেয়ে দেখে... একদম হার্ড লাগবে না..!!
আজ ভারত কে হারিয়ে ছাড়ব
অভদ্র মিমস
রেডি হয়ে মাঠে নাম।
আজ রেন্ডি নাচ নাচাব

Label: Hate

Hate Cateories: Polticial

Targeted Groups: Organization

Text[Ban]: যখন তুমি বিএনপির জন্য প্রচার-প্রচারনা চালিয়ে আওয়ামী লীগকে ভোট দেও: এই শহরে আমার মতো ক্রিমিনাল আর একটাও নেই।
Text[Eng]: When you campaign for BNP but vote for Awami League, there is no bigger criminal in this city than me.

Label: Hate

Hate Cateories: Personal Offense

Targeted Groups: Individual

Text[Ban]: তোমার কি মন খারাপ? না আমার পেট, ফোন, কপাল, মুড, লাইফ সব খারাপ। চেহারা বলতে ভুলে গেছো?
Text[Eng]: Are you upset? Nope, it's my stomach, phone, forehead, mood, and whole life that are a mess. Did you forget to mention your face?

Label: Hate

Hate Cateories: Gender

Targeted Groups: Community

Text[Ban]: যখন মেয়েরা দেখতে পায় তার মতো একই ড্রেস পরে আর একটা মেয়ে এসেছে এবং তাকে ওই ড্রেসে বেশি রোগা লাগছে। ঢং দেখে মরে যাই।
Text[Eng]: When girls see another girl wearing the same dress a as them and she looks sickly in it. I die at the attitude.

Label: Hate

Hate Cateories: Religious

Targeted Groups: Community

Text[Ban]: বাবু বেশি খারাপ লাগলে একটু পানি খেয়ে নাও, রোজা ভাঙ্গবেনা।
Text[Eng]: Baby, if you feel too bad, have a little water, it won't break your fast.

Label: Non Hate

Text[Ban]: ভাই ইংলিশ এতো হার্ড কেনো? জল মিশিয়ে খেয়ে দেখে, একদম হার্ড লাগবে না..!!
Text[Eng]: Bro, why is English so hard? Mix it with water and try eating, it won't feel hard at all..!!

Label: Hate

Hate Cateories: Abusive

Targeted Groups: Society

Text[Ban]: আজ ভারত কে হারিয়ে ছাড়ব। রেডি হয়ে মাঠে নাম। আজ রেন্ডি নাচ নাচাব।
Text[Eng]: Today we'll beat India. Get ready to take the field. Today we'll make them dance like whores.

Figure 6: Examples of annotated memes from the BANHATEME dataset, covering Non-Hate and hateful memes across the Abusive, Political, Gender, Personal Offense, and Religious categories, with their corresponding targeted group annotations.

basis of religious belief, practice, or disbelief. Such content often frames religion as inferior or dangerous to justify exclusion.

Indicators:

- Mockery of religious practices, figures, or rituals
- Advocacy of discrimination or exclusion of religious minorities
- Depicting a religion or its followers as violent, corrupt, or inferior

A.3 Targeted Groups and Definitions

- **Individual:** Hate directed at a specific person, often exploiting characteristics such as gender, popularity, race, or social standing. Targets may include both public figures and private individuals.
- **Organization:** Hate targeting an established body or institution composed of multiple people working toward shared goals. This includes corporations, educational or governmental institutions, and political parties.
- **Community:** Hate directed toward a collective of individuals who share common beliefs, practices, or affiliations. Such groups may be defined by religion, cultural tradition, fandom, or political alignment.
- **Society:** Hate generalized toward a broad population defined by nationality, ethnicity, or geography. This category captures content that vilifies entire societies or nations rather than a single group or organization.

B Annotation Tool

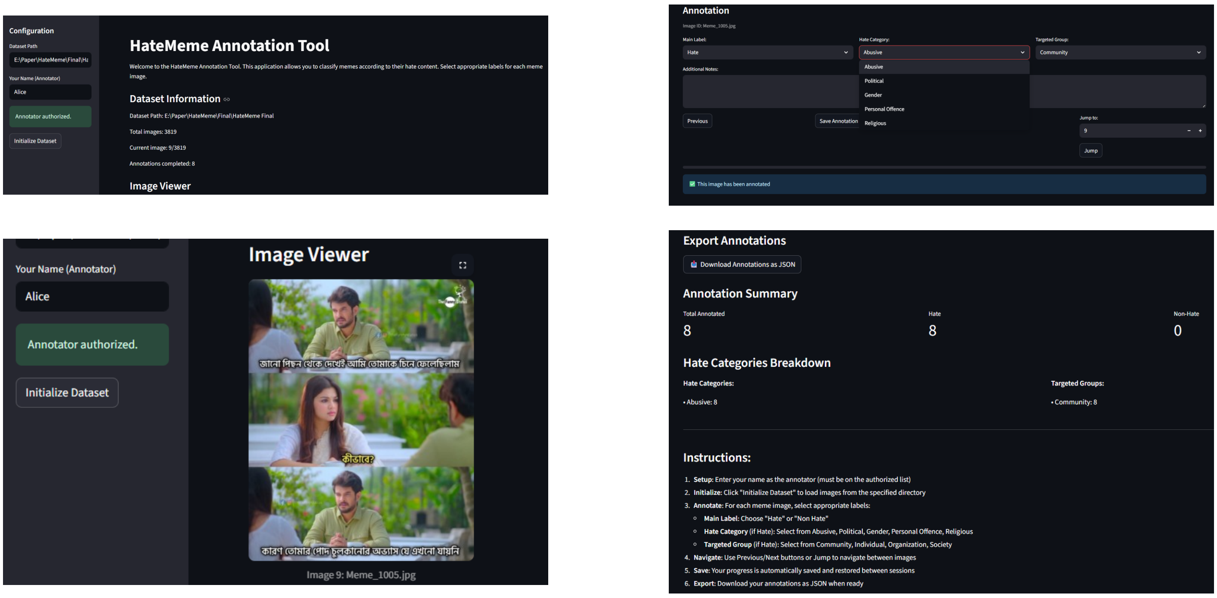
To ensure systematic and reliable labeling, we developed a dedicated web-based annotation tool customized for the BANHATEME dataset. The interface integrates the full hierarchical annotation process and was designed to reduce annotator workload while maintaining consistency across samples. Figure 7 illustrates the user interface, which incorporates several key components to support efficient multi-level annotation.

The platform provides the following core functionalities:

- **Authentication and Setup:** The left sidebar provides secure login, dataset path configuration, and annotation initialization. Access control ensures that only authorized annotators can provide labels.
- **Image Display:** Each meme is shown at its original resolution in the central panel, enabling annotators to examine both visual and textual elements in context.
- **Main Label Selection:** A drop-down menu requires annotators to assign a binary label (Hate vs. Non-Hate). If Non-Hate is chosen, the subsequent menus for hate category and targeted group are automatically set to N/A, preventing mislabeling.
- **Hierarchical Classification (if Hate):**
  - **Hate Category:** A structured drop-down menu enforces single selection from five predefined categories (Abusive, Political, Gender, Personal, or Religious).
  - **Targeted Group:** A second drop-down menu requires annotators to select exactly one of four group types (Community, Individual, Organization, or Society).
- **Contextual Notes:** A free-text field allows annotators to record observations, cultural cues, or rationale for their decisions.
- **Progress Management:** Navigation controls (Previous, Next, Jump) and automatic progress saving facilitate uninterrupted annotation and accurate tracking.
- **Data Export:** Annotations can be exported directly in JSON format, ensuring compatibility with validation scripts and downstream machine learning workflows.

C Error Analysis

Qualitative Analysis on Fusion Strategies. Figure 9 shows how different fusion strategies perform under the BanglaBERT+Swin configuration. Concatenation emerged as the most consistent, particularly in cases where religious or political hate was directed toward communities; by preserving modality-specific signals, it successfully aligned explicit textual references with symbolic visual cues. In contrast, summation often weakened discriminative information, leading to errors in nuanced cases such as personal offence or political hate toward organizations, where subtle textual



HateMeme Annotation Tool

Welcome to the HateMeme Annotation Tool. This application allows you to classify memes according to their hate content. Select appropriate labels for each meme image.

Dataset Path: E:\Paper\HateMeme\Final\Ha

Your Name (Annotator): Alice

Annotator authorized.

Dataset Information

Dataset Path: E:\Paper\HateMeme\Final\HateMeme Final

Total images: 3819

Current image: 9/3819

Annotations completed: 8

Image 9: Meme_1005.jpg

Annotation

Image ID: Meme_1005.jpg

Main Label: Hate

Hate Category: Abusive

Targeted Group: Community

Additional Notes:

Jump to: 9

☑ This image has been annotated

Annotation Summary

Total Annotated	Hate	Non-Hate
8	8	0

Hate Categories Breakdown

Hate Categories:

- Abusive: 8

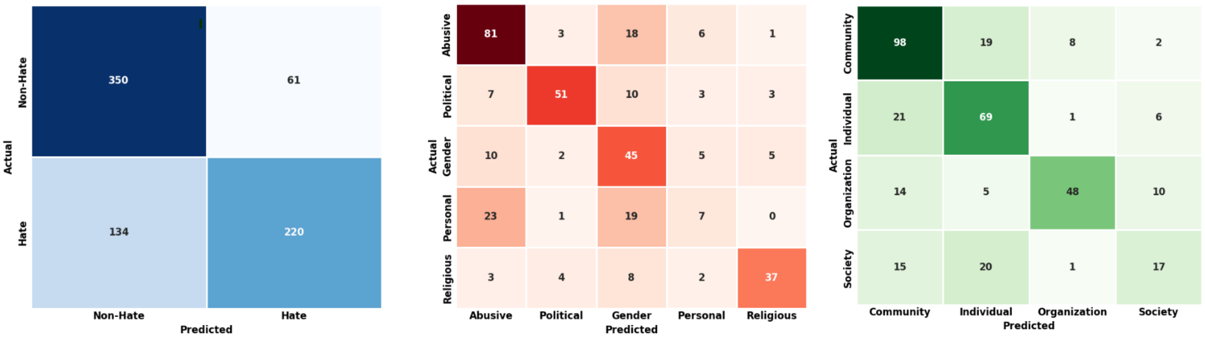
Targeted Groups:

- Community: 8

Instructions:

1. **Setup**: Enter your name as the annotator (must be on the authorized list)
2. **Initialize**: Click "Initialize Dataset" to load images from the specified directory
3. **Annotate**: For each meme image, select appropriate labels:
   - **Main Label**: Choose "Hate" or "Non Hate"
   - **Hate Category** (if Hate): Select from Abusive, Political, Gender, Personal Offence, Religious
   - **Targeted Group** (if Hate): Select from Community, Individual, Organization, Society
4. **Navigate**: Use Previous/Next buttons or Jump to navigate between images
5. **Save**: Your progress is automatically saved and restored between sessions
6. **Export**: Download your annotations as JSON when ready

Figure 7: Interface of the web-based annotation tool for the BanHateMe dataset, highlighting the configuration panel, annotation workspace, and meme display with hierarchical labeling functionality.



(a) Binary Label

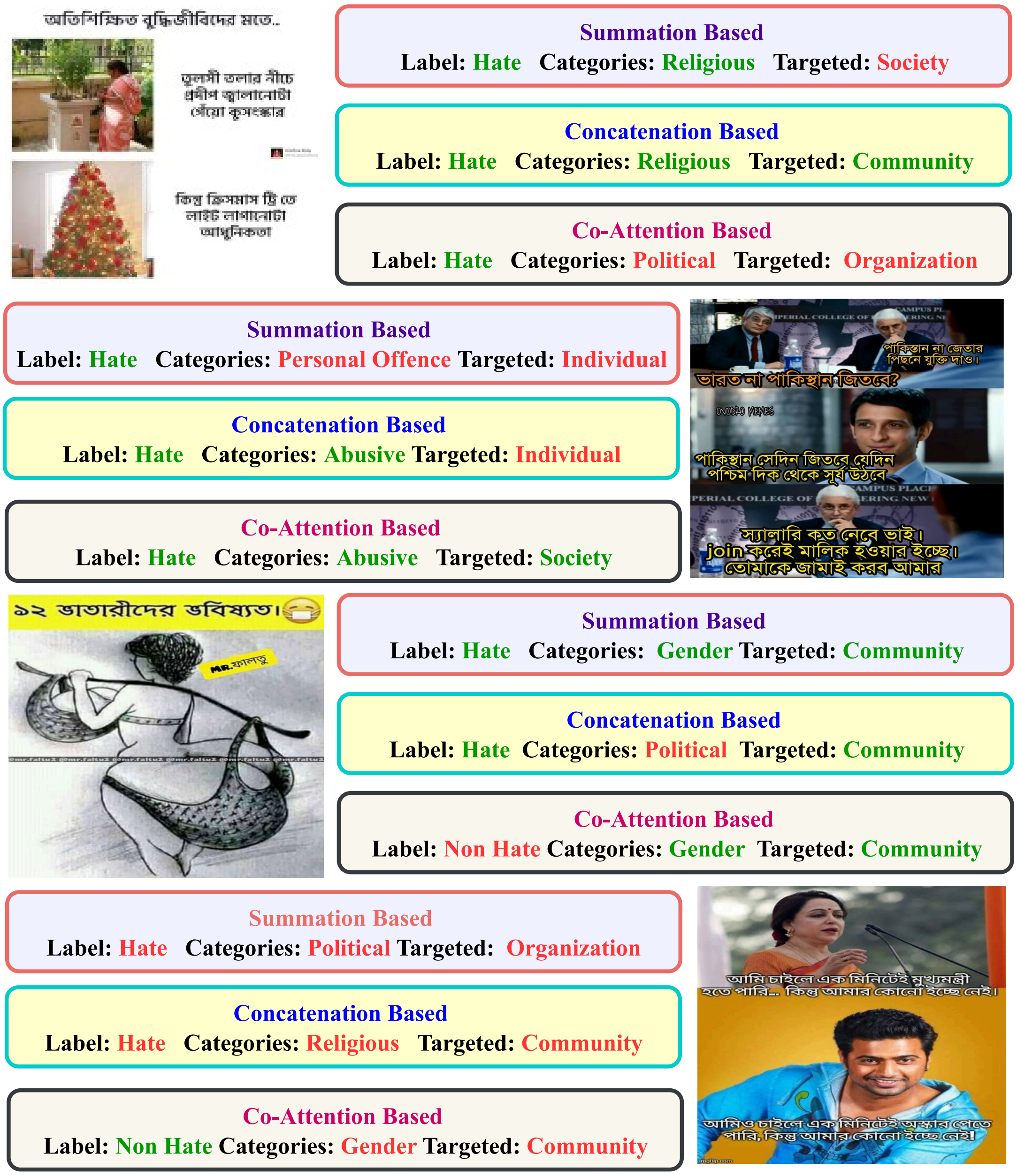
(b) Hate Categories

(c) Targeted Groups

Figure 8: Quantitative error analysis with BanglaBERT + Swin + Concatenation across binary, category, and target group levels.

phrasing or localized visual details were overshadowed by averaged features. Co-attention, while stronger at reasoning over target groups such as society, sometimes over-focuses on a single modality, causing failures like misclassifying offensive gendered caricatures as non-hate. Overall, concatenation proved most reliable for binary and categorical detection, co-attention offered complementary strengths for group-level inference but lacked stability, and summation consistently lagged behind.

Quantitative Error Analysis. We analyze errors for the BanglaBERT + Swin + Concatenation configuration, focusing on binary classification, hate categories, and target groups. In the binary task, as reported in Figure 8(a), nearly 40% of hateful memes are misclassified as non-hate, reflecting the challenge of detecting implicit or sarcastic expressions where cues are subtle. At the category level (Figure 8(b)), Abusive emerges as the most stable class, while Political and Gender often overlap with Abusive, showing leakage of about 15–20%. Personal Offence proves particularly difficult, with more than 30% of instances mislabeled as Abusive or Gender, while Religious hate is sometimes confused with Gender. For target groups (Figure 8(c)), Community is the most consistently recognized, while Individuals show moderate reliability but are redirected into Community about 20% of the time. Organization is less robust, with roughly 25% of its samples misclassified as Community, and Society is the most error-prone, with over 30% of instances mislabeled as either Community or Individual. Overall, these errors indicate that while binary detection is relatively strong, fine-grained categories and target groups remain substantially harder to distinguish due to overlapping linguistic signals and subtle visual markers.


অতিশিক্ষিত বুদ্ধিজীবিদের মতে..
তুলসী তলার নীচে প্রদীপ জ্বালানোটা গেঁয়ো কুসংস্কার
কিন্তু ক্রিসমাস ট্রি তে লাইট লাগানোটা আধুনিকতা

Summation Based
Label: Hate Categories: Religious Targeted: Society

Concatenation Based
Label: Hate Categories: Religious Targeted: Community

Co-Attention Based
Label: Hate Categories: Political Targeted: Organization

Summation Based
Label: Hate Categories: Personal Offence Targeted: Individual

Concatenation Based
Label: Hate Categories: Abusive Targeted: Individual

Co-Attention Based
Label: Hate Categories: Abusive Targeted: Society

পাকিস্তান না জেতার পিছনে যুক্তি দাও।
ভারত না পাকিস্থান জিতবে?
OVODRO MEMES
পাকিস্থান সেদিন জিতবে যেদিন পশ্চিম দিক থেকে সূর্য উঠবে
স্যালারি কত নেবে ভাই। join করেই মালিক হওয়ার ইচ্ছে। তোমাকে জামাই করব আমার

৯২ ভাতারীদের ভবিষ্যত।😷
MR.ফালতু
@mr.faltu2 @mr.faltu2 @mr.faltu2 @mr.faltu2 @mr.faltu2 @mr.faltu2

Summation Based
Label: Hate Categories: Gender Targeted: Community

Concatenation Based
Label: Hate Categories: Political Targeted: Community

Co-Attention Based
Label: Non Hate Categories: Gender Targeted: Community

Summation Based
Label: Hate Categories: Political Targeted: Organization

Concatenation Based
Label: Hate Categories: Religious Targeted: Community

Co-Attention Based
Label: Non Hate Categories: Gender Targeted: Community

আমি চাইলে এক মিনিটেই মুখ্যমন্ত্রী হতে পারি... কিন্তু আমার কোনো ইচ্ছে নেই।
আমিও চাইলে এক মিনিটেই অস্কার পেতে পারি, কিন্তু আমার কোনো ইচ্ছে নেই!
imgflip.com

Figure 9: Qualitative error analysis of fusion strategies under the BanglaBERT+Swin configuration. Green indicates correct predictions, while red indicates errors.