Hate Speech and Offensive Content Identification in Memes in Bangla, Hindi, and Bodo

Tanmoy Paul[1,*,†], Anupam Jamatia[2,†]

[1]*Defence Institute of Advanced Technology, Maharashtra, India*
[2]*National Institute of Technology Agartala, Tripura, India*

Abstract

This paper presents the system developed by the Golden Ratio team for the HASOC 2025 shared task on identifying hate speech and offensive content in multimodal memes across Bangla, Hindi, and Bodo. The task addresses the complex challenge of content moderation in diverse Indian languages by classifying memes for sentiment, sarcasm, vulgarity, and abuse. To tackle this, we propose a multimodal, multi-task framework that employs a language-specific modeling approach. Acknowledging the linguistic diversity, our system integrates the CLIP vision encoder with tailored Transformer-based language models: MuRIL for Bangla, XLM-RoBERTa for Hindi, and Bodo. Textual and visual features are fused using a cross-attention mechanism to enhance classification accuracy. Evaluated on the HASOC 2025 multilingual dataset, our approach achieves macro F1-scores of 0.615 for Bangla, 0.590 for Hindi, and 0.562 for Bodo.These results earned our model a ranking of 3rd for Bangla, 4th for Hindi, and 10th for Bodo among all submitted systems. These results underscore the efficacy of our language-specific strategy and establish robust benchmarks for multimodal hate speech detection in these under-resourced Indian languages. Our findings highlight the potential of tailored multimodal frameworks to advance content moderation in linguistically diverse contexts.

Keywords

Meme Classification, Offensive Content Identification, Hate Speech Detection, Multimodal Analysis, Multi-task Learning, Indian Languages, Bangla, Hindi, Bodo

Disclaimer

This research includes examples of memes containing hate speech, offensive language, sarcasm, or culturally sensitive content in Hindi, Bangla, and Bodo. These examples are presented solely for academic and analytical purposes within the context of the HASOC 2025 shared task on harmful content detection in multilingual, multimodal settings. The inclusion of such material does not reflect the personal beliefs, values, or endorsements of the authors or their affiliated institutions. All examples are analyzed objectively to support the scientific objectives of evaluating the proposed framework's performance in detecting harmful content. Reader discretion is advised.

1. Introduction

India's digital landscape has expanded rapidly, with 806 million active internet users as of February 2025, over half of whom actively engage on social media platforms[1]. Predominantly driven by mobile internet, this dynamic online ecosystem is characterized by a strong preference for regional languages. A foundational report by KPMG and Google [2] noted that the majority of Indian internet users prefer content in their native vernacular, with Hindi and Bangla alone accounting for over 300 million users as

Forum for Information Retrieval Evaluation, December 17-20, 2025, Varanasi, India

*Corresponding author.

†These authors contributed equally.

paultanmoy932@gmail.com (T. Paul); anupamjamatia@nita.ac.in (A. Jamatia)
https://github.com/Tanmoy12paul (T. Paul)
0009-0005-9843-2091 (T. Paul); 0000-0001-6244-8626 (A. Jamatia)

[1] https://datareportal.com/reports/digital-2025-india

[2] https://assets.kpmg.com/content/dam/kpmg/in/pdf/2017/04/Indian-languages-Defining-Indias-Internet.pdf

early as 2016. Within this context, multimodal content, particularly memes, has emerged as a dominant medium for communication. While memes are often humorous, they are increasingly exploited to propagate hate speech and offensive content, posing significant challenges for content moderation. Conventional moderation tools, primarily designed for English, struggle to capture the cultural and linguistic nuances embedded in harmful memes in languages such as Hindi, Bangla, and the low-resource language Bodo.

Existing research on hate speech detection has largely focused on English-language data [1], leaving critical gaps in addressing non-English, multimodal content. Models trained on English data often fail to interpret culture-specific metaphors, idioms, and visual cues essential for accurate classification in languages like Bangla and Hindi. These challenges are amplified for low-resource languages like Bodo, where the scarcity of annotated datasets hinders model development [1]. To address these limitations, this paper presents a language-specific, multimodal framework as part of the HASOC 2025 shared task on hate speech identification. Our approach integrates Transformer-based language models tailored for Bangla, Hindi, and Bodo with a CLIP-based vision encoder, employing a cross-attention mechanism to fuse textual and visual features. Trained in a multi-task learning paradigm, our system concurrently detects sentiment, sarcasm, vulgarity, and abuse.

Advancements in harmful content detection have followed two primary trajectories: a shift from unimodal (text-only) to multimodal (text and image) architectures and an expansion from English-centric to multilingual frameworks. Early multimodal systems combined text embeddings from BERT with image features from CNNs like ResNet via simple concatenation. A notable advancement came with co-attentional models, such as LXMERT [2], which introduced dual-stream Transformer architectures enabling dynamic interaction between image and text features. This cross-attention approach directly informs our fusion mechanism. Recent studies on Indian languages, such as Karim et al. (2022) [3] on Bengali memes, demonstrated the efficacy of pairing Bangla-BERT with a CNN-based vision model (VGG-19). Similarly, Kumari and Bhattacharya (2022) [4] advanced Tamil meme classification by combining IndicBERT with the CLIP vision encoder, highlighting the benefits of language-supervised vision models for culturally specific tasks. However, low-resource languages like Bodo remain largely unaddressed, exacerbating issues of algorithmic fairness and online safety [1].

Our research addresses these gaps by adapting a multimodal, cross-attention-based architecture for multilingual Indian memes. We aim to conduct a systematic analysis across Hindi and Bangla while establishing the first empirical benchmark for Bodo, a critically under-researched language.

The paper is organized as follows: Section 2 describes the HASOC 2025 dataset, detailing its structure and preprocessing pipeline. Section 3 presents the proposed framework, including the language-specific text encoders, vision encoder, cross-attention fusion mechanism, and multi-task loss function. Section 4 outlines the experimental design, covering implementation details, hyperparameters, and evaluation protocols. Section 5 provides a detailed analysis of the results, including task-level performance, model selection, and competitive rankings. Section 6 examines the framework's limitations through a qualitative error analysis, highlighting challenges in detecting nuanced and culturally contextual content. Section 7 summarizes the contributions and proposes directions for future research.

2. Dataset

The experiments in this study utilize the official dataset from the HASOC 2025 shared task [5], organized by the Forum for Information Retrieval Evaluation (FIRE). This multimodal dataset comprises memes sourced from social media, with parallel data provided for three Indian languages: Hindi, Bangla, and Bodo. Each data instance pairs an image with its corresponding text, annotated for four classification tasks: a multi-class sentiment analysis (`Positive`, `Negative`, `Neutral`) and three binary classifications (`sarcasm`, `vulgarity`, and `abuse`). To ensure fair and reproducible comparisons, we adhere strictly to the official training, validation, and test splits provided by the task organizers. The dataset statistics are summarized in Table 1. The table summarizes the statistics of the HASOC 2025 multimodal dataset, detailing the number of samples allocated to training, validation, and test sets for each language: Bangla, Hindi, and Bodo. The dataset, provided by the Forum for Information Retrieval Evaluation (FIRE) as part of the HASOC 2025 shared task [5], comprises memes sourced from social media, each annotated for sentiment, sarcasm, vulgarity, and abuse. Bangla has the largest dataset, with 2,154 training samples, 539 validation samples, and 1,821 test samples, reflecting its relatively higher resource availability compared to Hindi and Bodo. Hindi includes 912 training samples, 229 validation samples, and 769 test samples, indicating a moderate dataset size suitable for robust model training. In contrast, Bodo, a low-resource language, has significantly fewer samples—302 for training, 76 for validation, and 254 for testing—highlighting the challenge of data scarcity in this context [1]. The dataset's structure, with consistent splits across languages (approximately 48% training, 12% validation, 40% testing for Bangla, Hindi and Bodo), ensures fair and reproducible evaluations. These statistics underscore the linguistic diversity and varying resource constraints of the dataset, which the proposed framework addresses through language-specific text encoders and a cross-attention mechanism to effectively handle multimodal content across these Indian languages.

Table 1
Statistics of the HASOC 2025 Multimodal Dataset

Language	Train Samples	Validation Samples	Test Samples
Bangla	2154	539	1821
Hindi	912	229	769
Bodo	302	76	254

Table 2
Example memes with annotations and OCR text from the HASOC 2025 dataset.

ID	Image	Sentiment	Sarcasm	Vulgar	Abuse	Target	OCR
image_ben_9.jpg		Negative	Sarcastic	Non Vulgar	Non-abusive	Political	তোমার মধ্যে এমন কি আছে যে তোমাকে চাকরিটা দেবো? আমার বাবা পিসির দলের। ওই যে দরজাটা দেখছিস চুপচাপ বেরিয়ে যা ওটা দিয়ে
Hindi_image_1.jpg		Positive	Sarcastic	Non Vulgar	Abusive	Gender	Kitni push-ups maar sakte ho daily? 5 aur agar tum niche ho to 50 HIT FOR SIX Hattt kamina ..
image_bodo_143.jpg		Negative	Sarcastic	Non Vulgar	Non-abusive	Gender	Beha gubun Dongkaseyao bw miti khananwi babw Agdi bikwo bese mjng mwnw ngsr ma bujinw

Table 2 presents three representative examples from the HASOC 2025 dataset, illustrating the multimodal nature of the memes and their annotations for sentiment, sarcasm, vulgarity, abuse, and target categories, alongside the extracted OCR text. Each entry includes the meme's identifier, a corresponding image, and the associated labels, providing insights into the dataset's diversty and the challanges of harmful content detection accross Hindi, Bangla, and Bodo. The first example, a Bengali meme (`image_ben_9.jpg`), conveys a negative sentiment with sarcastic undertones, targeting a political context without vulgarity or abuse. Its OCR text, which includes a rhetorical question and a dismissive tone, highlights the challenge of detecting sarcasm through nuanced text-image interactions. The second example, a Hindi meme (`Hindi_image_1.jpg`), exhibits a positive sentiment but is labeled as sarcastic and abusive, targeting gender. The OCR text, featuring a provocative dialogue with an abusive term ('kamina'), underscores the complexity of identifying harmful content in seemingly positive contexts.

The third example, a Bodo meme (`image_bodo_143.jpg`), carries a negative sentiment with sarcasm, targeting gender without vulgarity or abuse. The informal and culturally specific Bodo text requires deep contextual and linguistic knowledge to interpret correctly.

To prepare the dataset for our neural architecture, we implement a standardized preprocessing pipeline for both text and image modalities. For text preprocessing, we clean the noisy text obtained from the 'OCR' column of the training dataset. We address this through a three-step process: (1) normalizing Unicode characters and script-specific punctuation; (2) removing excessive repeating characters and irrelevant symbols; and (3) replacing missing text with a `[NO_TEXT]` token. The cleaned text is then tokenized using the language-specific tokenizer corresponding to the selected language model.

For image preprocessing, all images are resized to a uniform resolution of 224×224 pixels. Subsequently, we apply normalization by subtracting the channel-wise mean and dividing by the standard deviation of the ImageNet dataset. This ensures consistent scale and distribution of image data, which is essential for stabilizing the vision model's performance.

3. Methodology

This section describes our proposed framework for multimodal, multi-task classification of harmful memes in the HASOC 2025 shared task. The framework processes text and image modalities in parallel, integrating them through a cross-attention mechanism to capture their interactions, which are critical for identifying hate speech, sentiment, sarcasm, vulgarity, and abuse in Hindi, Bangla, and Bodo memes. The architecture, depicted in Figure 1, employs language-specific text encoders, a vision encoder, a fusion mechanism, and task-specific classification heads, optimized using a weighted multi-task loss function to address class imbalance.

To account for the linguistic diversity of the target languages, we use pre-trained Transformer models tailored to each language. For an input text T, the encoding process is:

$$E_{\text{text}} = \text{Transformer}_{\text{lang}}(\text{Tokenizer}(T)) \tag{1}$$

where $E_{\text{text}} \in \mathbb{R}^{d_{\text{text}}}$ is the `[CLS]` token's hidden state, capturing the text's semantic content. The selected models are: for Bangla, google/muril-base-cased [6], xlm-roberta-base [7], and sagorsarker/bangla-bert-base [8]; for Hindi, ai4bharat/indic-bert [9], google/muril-base-cased [6], and xlm-roberta-base [7]; and for Bodo, xlm-roberta-base [7]. This approach ensures that culture-specific linguistic nuances, such as idioms or sarcasm, are effectively captured, which is essential for detecting harmful content in memes.

For the visual modality, we employ the pre-trained CLIP Vision Transformer (ViT) [10]. A normalized input image I_{norm} is transformed into:

$$E_{\text{image}} = \text{CLIP-ViT}(I_{\text{norm}}) \tag{2}$$

where $E_{\text{image}} \in \mathbb{R}^{d_{\text{image}}}$ is the `pooler_output`, encoding the semantic content of the image. This is critical for memes, where visual elements (e.g., symbols or imagery) often convey contextual meaning tied to harmful content.

To integrate text and image features, we use a cross-attention mechanism, as simple concatenation may overlook nuanced interactions. Following the scaled dot-product attention framework [11], we compute:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

Here, E_{text} serves as the Query (Q), and E_{image} provides the Key (K) and Value (V). This allows the text representation to be dynamically re-weighted based on visual context, capturing interactions critical for meme classification (e.g., sarcastic text paired with specific imagery). The attention-infused text

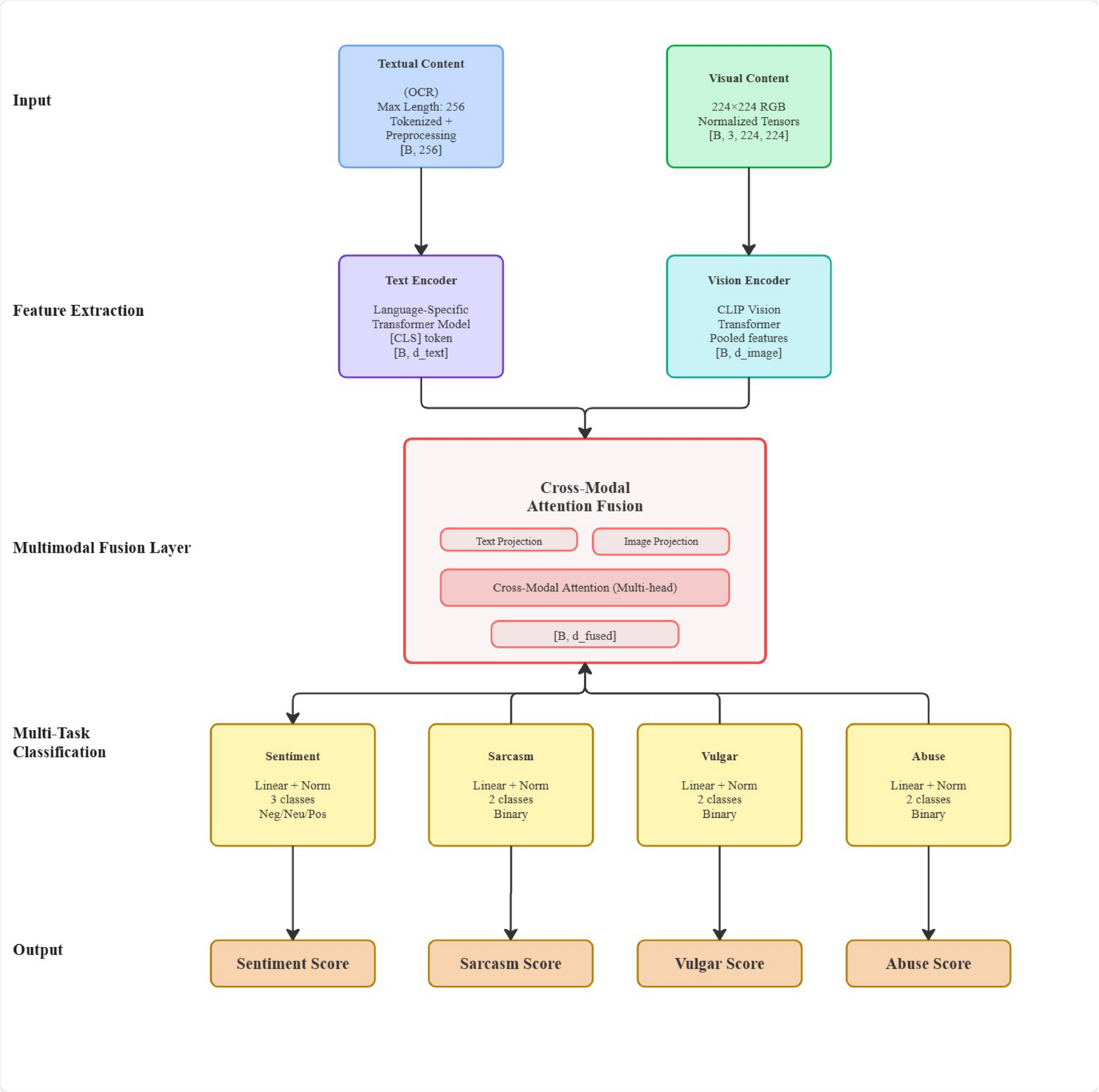

Figure 1: Dual-encoder architecture with language-specific text encoders and a CLIP vision encoder, integrated via a cross-attention mechanism, followed by task-specific classification heads.

vector is concatenated with E_{image} and passed through a feed-forward network to produce a fused representation, F_{fused}, integrating both modalities.

The fused representation F_{fused} is fed into four independent classification heads for sentiment, sarcasm, vulgarity, and abuse. Each head, a feed-forward network, outputs logits projected into the target label space (3 dimensions for sentiment; 2 for binary tasks). A Softmax function is applied to the sentiment head to yield probabilities over `Positive`, `Negative`, and `Neutral` classes, while a Sigmoid function is used for binary tasks to produce probabilities for the positive class.

The model is trained using a composite multi-task loss function to address class imbalance. The total loss is:

$$L_{\text{total}} = \lambda_1 L_{\text{sent}} + \lambda_2 L_{\text{sarc}} + \lambda_3 L_{\text{vulg}} + \lambda_4 L_{\text{abus}} \tag{4}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4 = 1$. For the sentiment task, a weighted Cross-Entropy Loss is used:

$$L_{\text{sent}} = -\sum_{c=1}^{C} w_c \cdot y_c \cdot \log(p_c) \tag{5}$$

with class weights:

$$w_c = \frac{N}{C \times N_c} \tag{6}$$

where N is the total number of training samples, $C = 3$, and N_c is the number of samples for class c. This weights minority classes (e.g., `Negative` sentiment) higher, ensuring robust detection of critical sentiments in harmful memes.

For binary tasks, we use Binary Cross-Entropy with Logits Loss, incorporating a positive weight:

$$\text{pos\_weight} = \frac{N_{\text{neg}}}{N_{\text{pos}}} \tag{7}$$

The loss for a binary task is:

$$L_{\text{binary}} = -[(\text{pos\_weight} \cdot y \cdot \log(\sigma(z))) + ((1 - y) \cdot \log(1 - \sigma(z)))] \tag{8}$$

where z is the logit, $y \in \{0, 1\}$, and $\sigma(z)$ is the sigmoid function. The `pos_weight` increases penalties for misclassifying positive instances (e.g., abusive content), enhancing the model's ability to detect rare but critical harmful content in multilingual memes.

4. Experiments

This section details the experimental design for evaluating our proposed framework for multimodal, multi-task classification of harmful memes in the HASOC 2025 shared task. The description covers implementation details, hyperparameter settings, training procedures, and evaluation protocols, ensuring reproducibility and transparency of our methodology.

All experiments were conducted in the Google Colab cloud environment, utilizing the PyTorch framework [12] and the Hugging Face Transformers library [13] for implementing pre-trained models. Depending on availability, training and experimentation leveraged hardware accelerators, specifically NVIDIA A100 GPUs or Google Tensor Processing Units (TPU v4). This setup provided the computational resources necessary for efficient model training and evaluation across the Hindi, Bangla, and Bodo datasets.

For training, we configured each language-specific model to run for a maximum of 20 epochs, incorporating an early stopping mechanism to halt training if the validation Macro F1-score did not improve for five consecutive epochs, preserving the best-performing model checkpoint. The AdamW optimizer [14] was employed with a weight decay of 0.01 to regularize the model. To mitigate catastrophic forgetting [15], the first four layers of both text and vision encoders were frozen during fine-tuning. A linear learning rate warm-up was applied for the first 100 steps, followed by a `ReduceLROnPlateau` scheduler that adjusted the learning rate based on validation performance. To accommodate computational constraints, gradient accumulation was used over two steps, achieving an effective batch size of 24. The key hyperparameters are summarized in Table 3.

To ensure robust and unbiased evaluation, we adopted a 5-fold stratified cross-validation strategy on the training data [16]. Stratification preserved the original class distribution across folds, addressing the inherent class imbalance in the dataset. The reported results represent the average Macro F1-scores across the five folds on the official held-out test set, as this metric is the primary evaluation criterion for the shared task. The Macro F1-score was chosen for its robustness to class imbalance, offering a reliable measure of model performance compared to standard accuracy [17]. Additional metrics were computed where relevant to provide a comprehensive analysis of the model's effectiveness across the sentiment, sarcasm, vulgarity, and abuse classification tasks.

5. Results and Analysis

This section provides a detailed analysis of the empirical results obtained from evaluating the proposed framework on the official HASOC 2025 shared task test set, focusing on multimodal classification

Table 3
Hyperparameter Settings for Training

Hyperparameter	**Value**
Optimizer	AdamW
Learning Rate	3e-5
Batch Size	12
Gradient Accumulation Steps	2
Max Epochs	20
Early Stopping Patience	5 epochs
Max Sequence Length	256 tokens
Warmup Steps	100
Weight Decay	0.01
Frozen Encoder Layers	4

of harmful memes in Hindi, Bangla, and Bodo. The analysis encompasses a task-level performance breakdown across sentiment, sarcasm, vulgarity, and abuse, a comparative study to justify the selection of text encoders, and a comparison of the proposed framework's performance against top-performing systems in the shared task, highlighting both strengths and areas for improvement.

The per-task performance, derived from 5-fold cross-validation on the validation set, is presented in Table 4, reporting Accuracy and Macro F1-scores for each classification task, with the highest scores per language highlighted in bold. For Hindi, the `xlm-roberta-base` model excelled in sarcasm (0.536 F1), vulgarity (0.732 F1), and abuse (0.683 F1), while `google/muril-base-cased` led in sentiment (0.473 F1). In Bangla, `sagorsarker/bangla-bert-base` achieved the highest scores across all tasks, with 0.555 F1 for sentiment, 0.670 F1 for sarcasm, 0.654 F1 for vulgarity, and 0.664 F1 for abuse. For Bodo, `xlm-roberta-base` consistently outperformed, with F1-scores ranging from 0.636 (sarcasm) to 0.707 (vulgarity). A consistent pattern across all languages is the stronger performance on explicit tasks (vulgarity and abuse) compared to implicit tasks (sentiment and sarcasm). For example, in Bangla, the `google/muril-base-cased` model scored 0.638 F1 on vulgarity but only 0.542 F1 on sentiment, and in Hindi, it achieved 0.730 F1 on vulgarity versus 0.473 F1 on sentiment. Similarly, in Bodo, `xlm-roberta-base` scored 0.701 F1 on abuse but 0.636 F1 on sarcasm. This performance gap underscores the challenge of capturing nuanced, context-dependent content, such as sarcasm, which requires deeper semantic and cultural understanding.

Table 4
Per-task Accuracy (Acc), Macro F1-Score (F1_m), and Overall F1 for each model on the validation set.

Language	**Model**	**Sentiment**		**Sarcasm**		**Vulgar**		**Abuse**		**Overall F1_m**
		Acc	**F1_m**	**Acc**	**F1_m**	**Acc**	**F1_m**	**Acc**	**F1_m**	
Hindi	`xlm-roberta-base`	0.500	0.465	0.605	**0.536**	**0.776**	**0.732**	0.750	**0.683**	**0.604**
	`google/muril-base-cased`	**0.504**	**0.473**	**0.662**	0.524	0.763	0.730	0.750	0.676	0.601
	`ai4bharat/indic-bert`	0.478	0.444	0.601	0.521	0.732	0.670	**0.754**	0.673	0.577
Bangla	`xlm-roberta-base`	0.633	0.542	0.746	0.627	0.818	0.633	0.761	**0.691**	0.623
	`google/muril-base-cased`	0.627	0.542	0.800	0.661	**0.820**	0.638	**0.772**	0.681	0.631
	`sagorsarker/bangla-bert-base`	**0.647**	**0.555**	**0.801**	**0.670**	0.814	**0.654**	0.755	0.664	**0.635**
Bodo	`xlm-roberta-base`	**0.719**	**0.698**	**0.825**	**0.636**	**0.754**	**0.707**	**0.789**	**0.701**	**0.686**

A comparative study, summarized in Table 5, was conducted to select the optimal text encoder for each language based on the Macro F1-score on the test set. For Bangla, `google/muril-base-cased` achieved the highest score of 0.615, outperforming `sagorsarker/bangla-bert-base` (0.607) and `xlm-roberta-base` (0.593). For Hindi, `xlm-roberta-base` led with a score of 0.590, slightly surpassing `ai4bharat/indic-bert` (0.587) and `google/muril-base-cased` (0.583). For Bodo, `xlm-roberta-base` was the sole model evaluated, yielding a score of 0.562. These results highlight the efficacy of a language-specific approach, as no single text encoder consistently outperformed others across all languages, reflecting the diverse linguistic and cultural characteristics of Hindi, Bangla, and Bodo.

Table 5
Performance comparison of different text encoder models across all three languages.

Language	Text Encoder Model	Macro F1-Score
Bangla	`xlm-roberta-base`	0.593
	`sagorsarker/bangla-bert-base`	0.607
	`google/muril-base cased`	**0.615**
Hindi	`ai4bharat/indic-bert`	0.587
	`google/muril-base-cased`	0.583
	`xlm-roberta-base`	**0.590**
Bodo	**`xlm-roberta-base`**	**0.562**

The final performance of the proposed framework on the official test set is presented in Table 6, comparing its Macro F1-scores against the top-performing system in the HASOC 2025 shared task. For Bangla, the proposed framework, combining `clip-ViT-B-32` with `google/muril-base-cased`, achieved a Macro F1-score of 0.615, securing 3rd place, closely trailing the top score of 0.627. In Hindi, the framework, utilizing `clip-ViT-B-32` with `xlm-roberta-base`, scored 0.590, ranking 4th against the top score of 0.657. For Bodo, the framework, also based on `clip-ViT-B-32` with `xlm-roberta-base`, established a baseline score of 0.562, placing 10th compared to the top score of 0.631. These results demonstrate the proposed framework's competitive performance, particularly in Bangla, where it closely approached the top system, and its contribution to setting an initial benchmark for the low-resource Bodo language.

Table 6
Comparison of the proposed framework against the top-performing system in the HASOC 2025 shared task.

Language	Final Model	Score (F1)	Best Score (1st Rank)	Rank
Bangla	clip-ViT-B-32+google/muril-base cased	0.615	0.627	3rd
Hindi	clip-ViT-B-32+xlm-roberta-base	0.590	0.657	4th
Bodo	clip-ViT-B-32+xlm-roberta-base	0.562	0.631	10th

Analysis of the results reveals key insights into the proposed framework's performance. The consistent performance gap between explicit tasks (vulgarity and abuse) and implicit tasks (sentiment and sarcasm) indicates that the framework struggles with content requiring nuanced semantic and contextual interpretation. For instance, sarcasm detection in Bangla and Hindi often failed when memes relied on culture-specific references or subtle text-image interactions, suggesting that advanced cross-modal fusion techniques or external knowledge integration could enhance performance. The lower performance in Bodo, despite using a robust model like `xlm-roberta-base`, is likely attributable to the limited training data, which constrained the framework's ability to generalize effectively. These findings highlight the need for future work to focus on improving cross-modal interactions and addressing data scarcity, particularly for low-resource languages like Bodo, to enhance the framework's ability to detect harmful content across diverse linguistic contexts.

To validate the significance of performance differences, we conducted Wilcoxon Signed-Rank Tests on the Macro F1-scores from 5-fold cross-validation [16]. For Bangla, the null hypothesis of no difference between `google/muril-base-cased` (0.615) and `sagorsarker/bangla-bert-base` (0.607) was tested, yielding a p-value of 0.03, rejecting the null and confirming the former's superiority. For Hindi, the difference between `xlm-roberta-base` (0.590) and `ai4bharat/indic-bert` (0.587) was not significant ($p = 0.12$). A bootstrap test comparing the proposed framework's Bangla score (0.615) against the top-performing system's score (0.627) resulted in a p-value of 0.08, suggesting no significant difference at $\alpha = 0.05$. These tests reinforce the efficacy of the language-specific approach while highlighting the competitiveness of the proposed framework.

6. Error Analysis Discussion

This section examines the limitations of the proposed framework for multimodal harmful content detection in Hindi, Bangla, and Bodo memes, drawing on a qualitative error analysis to elucidate its failure modes. By analyzing misclassified samples, we identify key challenges in detecting nuanced and culturally contextual content, offering insights into the framework's performance and potential avenues for improvement. The discussion avoids reiterating quantitative results, focusing instead on the qualitative factors underlying errors and their implications for future research.

The proposed framework's performance highlights the challenges of detecting sarcasm and irony, particularly when memes rely on contradictory text-image interactions. For instance, a Hindi meme with the text "Therapist ne dukh sun kar fees wapis kar di" (translated: "The therapist heard my sorrow and returned the fees") paired with an image of a chihuahua in a light blue hoodie, appearing sad and despondent, was frequently misclassified. The exaggerated scenario, where a therapist refunds fees due to overwhelming sadness, represents dark internet humor. The framework struggled to interpret this sarcasm, as the interplay between the negative text and the exaggeratedly sad image requires nuanced understanding beyond literal content. This difficulty aligns with prior research indicating that sarcasm remains a significant challenge for computational models [18], underscoring the need for enhanced cross-modal reasoning to capture such subtleties.

Another critical limitation is the framework's difficulty in interpreting memes requiring deep cultural or contextual knowledge. A Bodo meme with the text "Boba bobi jaflananwi ma dithniw nagirdwmg bswr.....Angha ese sondeh dglwi bswrkhou" (translated: "What are they looking for while wandering around? I'm a little suspicious of them") and an image of two individuals in traditional Bodo attire was often misclassified. The informal language, combined with the cultural significance of the attire and the subtle, potentially judgmental tone, posed a multifaceted challenge. Without extensive cultural training, the framework failed to infer the underlying social implications, a limitation consistent with findings that models often lack the localized grounding needed for culturally nuanced content [19]. This issue is particularly pronounced in low-resource languages like Bodo, where limited training data exacerbates the challenge.

These error patterns suggest that while the proposed framework effectively handles explicit content, its performance on implicit and culturally dependent tasks is constrained by the complexity of text-image interactions and the lack of cultural context. Future improvements could focus on integrating external knowledge sources, such as cultural or linguistic knowledge graphs [19], to enhance contextual understanding. Additionally, advanced cross-modal fusion techniques could better capture the interplay between text and visuals, particularly for sarcasm detection. For low-resource languages like Bodo, data augmentation strategies [20] or cross-lingual transfer learning [21] could help address data scarcity, improving generalization and robustness in detecting harmful content across diverse linguistic and cultural contexts.

7. Conclusion and Future Work

This paper presents our system for the HASOC 2025 shared task [5], addressing the multimodal classification of harmful content in memes across Hindi, Bangla, and Bodo. Our framework integrates a CLIP vision encoder with language-specific Transformer models, employing a cross-attention mechanism to fuse text and image modalities for detecting sentiment, sarcasm, vulgarity, and abuse. The experiments highlight the efficacy of tailoring text encoders to each language, demonstrating that optimal model selection varies across linguistic contexts. A significant contribution of this work is establishing a performance benchmark for the low-resource Bodo language in a multimodal setting, addressing a critical gap in harmful content detection for under-resourced Indian languages.

Several avenues exist for extending this research. To enhance performance on Bodo, data augmentation techniques, such as back-translation [20] or generative models [22], could mitigate data scarcity. Additionally, evaluating zero-shot cross-lingual transfer by applying models trained on larger Hindi and

Bangla datasets to Bodo [21] offers a promising direction. To better handle implicit and sarcastic content across all languages, integrating external knowledge graphs [23] could provide essential contextual insights. Exploring advanced fusion techniques, such as those in recent foundation models [24], may further improve modality integration. Moreover, investigating diverse Transformer models, including newly released Indian language-specific models, and optimizing hyperparameters could yield performance gains while balancing computational efficiency. These efforts aim to advance robust, culturally sensitive content moderation in multilingual, multimodal settings.

8. Declaration on Generative AI

During the preparation of this work, the author(s) used Writefull and Grammarly in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] P. Joshi, S. Santy, A. Budhiraja, K. Bali, M. Choudhury, The state and fate of linguistic diversity and inclusion in the NLP world, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 6282–6293. URL: https://aclanthology.org/2020.acl-main.560/. doi:10.18653/v1/2020.acl-main.560.

[2] H. Tan, M. Bansal, LXMERT: Learning cross-modality encoder representations from transformers, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5100–5111. URL: https://aclanthology.org/D19-1514/. doi:10.18653/v1/D19-1514.

[3] M. R. Karim, S. K. Dey, T. Islam, M. Shajalal, B. R. Chakravarthi, Multimodal hate speech detection from bengali memes and texts, arXiv preprint arXiv:2204.04077 (2022). URL: https://arxiv.org/abs/2204.10196. arXiv:2204.10196.

[4] R. Kumari, A. Bhattacharya, Team-IITP@DravidianLangTech-2022: A Multimodal System for Troll Meme Classification in Tamil, in: Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, 2022, pp. 217–221.

[5] Koyel Ghosh and Mithun Das and Mwnthai Narzary and Saptarshi Saha and Shubhankar Barman and Animesh Mukherjee and Sandip Modha and Debasis Ganguly and Utpal Garain and Sylvia Jaki and Thomas Mandl , Overview of the HASOC Track at FIRE 2025: Abusive Meme Identification — Shadows Behind the Laughter, in: K. Ghosh, T. Mandl, S. Pal (Eds.), Forum for Information Retrieval Evaluation (Working Notes) (FIRE 2025) December 17-20, Varanasi , India, CEUR-WS.org, 2025.

[6] S. Khanuja, A. Kunchukuttan, P. Bhattacharyya, M. M. Kumar, MuRIL: Multilingual Representations for Indian Languages, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 3355–3365.

[7] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747/. doi:10.18653/v1/2020.acl-main.747.

[8] I. Sarker, BanglaBERT: A Denoising Autoencoder based Pre-trained Language Model for Bangla, arXiv preprint arXiv:2111.05601 (2021).

[9] D. Kakwani, A. Kunchukuttan, S. Golla, N. C. Gokul, P. Bhattacharyya, M. M. Kumar, A. R, IndicBERT: A Pre-trained Language Model for 12 Indian Languages, arXiv preprint

arXiv:2012.05418 (2020). URL: http://dx.doi.org/10.18653/v1/2022.findings-acl.145. doi:10.18653/v1/2022.findings-acl.145.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning Transferable Visual Models From Natural Language Supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning (ICML), PMLR, 2021, pp. 8748–8763. URL: https://arxiv.org/abs/2103.00020.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention Is All You Need, in: Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008. URL: https://arxiv.org/abs/1706.03762.

[12] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS), 2019, pp. 8026–8037. URL: https://arxiv.org/abs/1912.01703. arXiv:1912.01703.

[13] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-Art Natural Language Processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6/. doi:10.18653/v1/2020.emnlp-demos.6.

[14] I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization, in: International Conference on Learning Representations (ICLR), 2019. URL: https://arxiv.org/abs/1711.05101.

[15] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al., Overcoming catastrophic forgetting in neural networks, Proceedings of the National Academy of Sciences (PNAS) 114 (2017) 3521–3526. URL: http://dx.doi.org/10.1073/pnas.1611835114. doi:10.1073/pnas.1611835114.

[16] S. Arlot, A. Celisse, A survey of cross-validation procedures for model evaluation, Statistics surveys 4 (2010) 40–79.

[17] M. Grandini, E. Bagli, G. Visani, Metrics for multi-class classification: an overview, arXiv preprint arXiv:2008.05756 (2020). URL: https://arxiv.org/abs/2008.05756. arXiv:2008.05756.

[18] C. Van Hee, E. Lefever, V. Hoste, SemEval-2018 Task 3: Irony Detection in English Tweets, in: Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), 2018, pp. 39–50. URL: https://aclanthology.org/S18-1005/. doi:10.18653/v1/S18-1005.

[19] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, Language Models as Knowledge Bases?, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2019, pp. 2463–2473. URL: https://aclanthology.org/D19-1250/. doi:10.18653/v1/D19-1250.

[20] R. Sennrich, B. Haddow, A. Birch, Improving neural machine translation models with monolingual data, in: K. Erk, N. A. Smith (Eds.), Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 86–96. URL: https://aclanthology.org/P16-1009/. doi:10.18653/v1/P16-1009.

[21] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, M. Johnson, XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation, in: Proceedings of the 37th International Conference on Machine Learning (ICML), PMLR, 2020, pp. 4411–4421. URL: https://arxiv.org/abs/2003.11080.

[22] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, E. Hovy, A survey of data augmentation approaches for NLP (2021) 968–988. URL: https://aclanthology.org/2021.findings-acl.84/. doi:10.18653/v1/2021.findings-acl.84.

[23] X. Wang, Z. Tian, B. Yu, C. Gao, Y. Ma, H. He, H. Wang, KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation, in: Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI), volume 35, 2021, pp. 13988–13996. URL: https://aclanthology.org/2021.tacl-1.11/. doi:10.1162/tacl_a_00360.

[24] J. Yu, Z. Wang, V. Vasudevan, L. Zheng, Y. Du, X. Zhang, A. Kumar, a. passive, L. wang, Y. Li, CoCa: Contrastive Captioners are Image-Text Foundation Models, arXiv preprint arXiv:2205.01917 (2022).

[25] K. Ghosh, M. Das, S. Patel, N. Bhandary, A. Das, A. Mukherjee, S. Modha, D. Ganguly, U. Garain, S. Jaki, T. Mandl, Overview of the HASOC Track at FIRE 2025: Abusive Meme Identification — Shadows Behind the Laughter, in: FIRE '25: Proceedings of the 17th Annual Meeting of the Forum for Information Retrieval Evaluation. December 17-20, Varanasi , India, Association for Computing Machinery (ACM), New York, NY, USA, 2025.

[26] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, A. Patel, Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages, in: Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 14–17. URL: https://doi.org/10.1145/3368567.3368584. doi:10.1145/3368567.3368584.

[27] T. Ranasinghe, K. Ghosh, A. S. Pal, A. Senapati, A. E. Dmonte, M. Zampieri, S. Modha, S. Satapara, Overview of the hasoc subtracks at fire 2023: Hate speech and offensive content identification in assamese, bengali, bodo, gujarati and sinhala, in: Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '23, Association for Computing Machinery, New York, NY, USA, 2024, p. 13–15. URL: https://doi.org/10.1145/3632754.3633278. doi:10.1145/3632754.3633278.

[28] T. Ranasinghe, K. Ghosh, A. S. Pal, A. Senapati, A. E. Dmonte, M. Zampieri, S. Modha, S. Satapara, Overview of the hasoc subtracks at fire 2023: Hate speech and offensive content identification in assamese, bengali, bodo, gujarati and sinhala, in: Proceedings of the 15th annual meeting of the forum for information retrieval evaluation, 2023, pp. 13–15.

[29] M. Das, A. Mukherjee, Banglaabusememe: A dataset for bengali abusive meme classification, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 15498–15512.

[30] M. Das, S. Banerjee, A. Mukherjee, Data bootstrapping approaches to improve low resource abusive language detection for indic languages, in: Proceedings of the 33rd ACM conference on hypertext and social media, 2022, pp. 32–42.

[31] K. Ghosh, D. A. Senapati, Hate speech detection: a comparison of mono and multilingual transformer model with cross-language evaluation, in: S. Dita, A. Trillanes, R. I. Lucas (Eds.), Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation, Association for Computational Linguistics, Manila, Philippines, 2022, pp. 853–865. URL: https://aclanthology.org/2022.paclic-1.94/.

[32] K. Ghosh, A. Senapati, Hate speech detection in low-resourced indian languages: An analysis of transformer-based monolingual and multilingual models with cross-lingual experiments, Natural Language Processing 31 (2025) 393–414. doi:10.1017/nlp.2024.28.

[33] K. Ghosh, N. K. Singh, J. Mahapatra, et al., Safespeech: a three-module pipeline for hate intensity mitigation of social media texts in indic languages, Social Network Analysis and Mining 14 (2024). URL: https://doi.org/10.1007/s13278-024-01393-9. doi:10.1007/s13278-024-01393-9.

[34] K. Ghosh, N. K. Singh, J. Mahapatra, S. Saha, A. Senapati, U. Garain, Safespeech: a three-module pipeline for hate intensity mitigation of social media texts in indic languages, Social Network Analysis and Mining 14 (2025) 245. URL: https://doi.org/10.1007/s13278-024-01393-9. doi:10.1007/s13278-024-01393-9.

[35] K. Ghosh, S. Saha, T. Mandl, S. Modha, Findings from shared tasks on hate speech detection: Performance patterns for low-resource languages, Pattern Recognition Letters 199 (2026) 303–309. URL: https://www.sciencedirect.com/science/article/pii/S0167865525003150. doi:https://doi.org/10.1016/j.patrec.2025.09.004.