A Multi-Modal Ensemble Approach for Hate Speech and Offensive Content Detection in Indic Memes

Radhika Bohra[1,*,†], Yashvardhan Sharma[1,†]

[1]Department of CSIS, Birla Institute of Technology and Science, Pilani, 333031, Rajasthan, INDIA

Abstract

The proliferation of multi-modal content, such as memes, on social media has created significant challenges for automated hate speech and offensive content detection, particularly in under-represented Indic languages. This paper presents a robust pipeline to address this challenge across four languages: Bangla, Bodo, Hindi, and Gujarati. We evaluate multiple dual-encoder architectures, combining the CLIP Vision Transformer with language-specific text models including MuRIL, XLM-Roberta, and M-BERT. Training is conducted using a 5-fold cross-validation strategy with a weighted loss function to counteract class imbalance, and performance is validated on the test set. Our results establish trustworthy performance benchmarks, with macro F1 scores consistently ranging from 0.56 to 0.61 across the different languages. A task-level analysis reveals that the models are effective at classifying sentiment, while more nuanced and context-dependent tasks like sarcasm remain challenging for current architectures. By systematically evaluating these strong multimodal models, this work provides a foundational benchmark that will guide and accelerate future research in content moderation for Indic languages. The research is conducted by the team of CSIS BITS Pilani. The team achieved the following ranks for the HASOC tasks on the 4 language datasets: Bangla - Rank 2, Bodo - Rank 5, Gujarati - Rank 6, and Hindi - Rank 8.

Keywords

Multi-modal, Hate speech and offensive content detection, Vision Language Models, Indic memes

1. Introduction

In little more than a decade, social media has transformed from a novelty into the world's digital gathering place. It's the digital equivalent of a nation-wide "chai stall", a bustling space where news breaks, opinions are forged, and an ever-growing volume of user-generated content is shared every second. Central to this digital culture is the rise of the meme. More than just simple jokes, memes are potent capsules of cultural shorthand, capable of conveying complex ideas, emotions, and commentary almost instantly, all through just an image and some text in it.

But this digital gathering place has a dark side. The very features that make memes effective for communication also make them an ideal vehicle for spreading hate speech and offensive content [1]. Hateful ideologies can be laundered through the use of irony or humour, packaged in a shareable format that makes them more palatable and viral. An otherwise innocent image can be combined with a subtle line of text to create a deeply abusive or derogatory message, which can negatively influence the public's opinion based on gender, religion, individual, politics, cultural identity, etc., while flying under the radar of traditional content moderation.

The rapid, unchecked spread of such material poses a direct threat to the safety of online communities [2, 3]. This has created an urgent need for effective detection, but identifying this content at scale is a formidable challenge. The core challenge in detecting offensive memes lies in their inherently multi-modal nature . The true intent of a meme is rarely found in the text or the image alone; instead, it emerges from the complex interplay between the two. The text might be benign, and the image harmless, but their combination can produce a potent and unambiguously hateful message. This

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

*Corresponding author.

†These authors contributed equally.

p20240428@pilani.bits-pilani.ac.in (R. Bohra); yash@pilani.bits-pilani.ac.in (Y. Sharma)

0009-0002-3584-1064 (R. Bohra)

semantic gap is often filled by implicit cultural, social, or political context that a machine must learn to infer. Furthermore, bad actors deliberately exploit this complexity, using sarcasm, irony, and coded language to evade traditional, text-based content moderation filters. This means any effective detection system must not only process both modalities but also understand the nuanced, often non-literal, relationship between them. Furthermore, the sheer scale and speed at which memes are created make manual moderation impossible. To foster healthier digital spaces for all users, we need automated systems sophisticated enough to understand this new, complex, and multilingual language.

To address this complex fusion of visual and textual data, the research community has increasingly turned to Vision-Language Models (VLMs). While current research is effectively focused on detecting hate and offensive content in English memes, one aspect being undermined in this field is the same challenge present in other prevalent languages across the globe.

This challenge is particularly acute for Indic languages, which are often critically under-represented in digital safety research despite their massive global presence. For example, Hindi is the third most spoken language in the world with more than 600 million speakers. Also, Bengali is the sixth most spoken native language and the seventh most spoken language in the world, with well over 270 million speakers. Yet, the development of robust moderation tools for Hindi, Bengali, and other languages of the subcontinent lags significantly behind that of English. This disparity creates a gap where harmful content can flourish, affecting millions of users.

Therefore, our research attempts to address this challenge. This paper tackles the task of multi-modal, multi-lingual, and multi-task classification of potential hate speech and offensive content present in memes in Indic languages. The goal of this research is to develop a robust pipeline to analyze Indic memes and classify them across four distinct categories: Sentiment, Sarcasm, Vulgarity, and Abuse. Our focus is on four Indic languages - Bodo, Bangla [4], Gujarati, and Hindi - to address the critical need for better tools in non-English contexts, accounting for challenges like code-mixed text ("Hinglish"). Furthermore, we address the difficulty of distinguishing between related but separate concepts, particularly Sarcasm, a nuanced form of expression that often complicates the detection of genuine hate speech.

The data for conducting this research was obtained from HASOC-meme 2025 [5, 6, 7, 8, 9, 10], a part of the Forum for Information Retrieval 2025. The data provided was a mix of images and their corresponding labels and text recognized by an OCR model. The presented research is an investigation into the limits of current state-of-the-art models on this complex-real world data. Our aim is to establish a trustworthy performance benchmark through a rigorous evaluation methodology and identify aspects on which future research can be done - whether it be in the model's architecture or in the data.

The primary contributions of this research are as follows:

- A VLM pipeline is proposed in this work that uses the Vision Transformer (ViT) of the CLIP VLM [11] for image processing and a comparison of three models - MuRIL [12], XLM-RoBERTa [13], and mBERT [14] - for the text processing. Finally the outputs of both are combined using a fusion layer and multi-task classification is performed for each image.
- Extensive experimental evaluation is conducted using the macro F1 Score as the primary performance metric to evaluate the performance of each of the models which provided a trustworthy performance benchmark.

The subsequent sections of this paper are organized as follows. The relevant literature related this research is reviewed in Section 2. The proposed methodology of this research is presented in Section 3. The results obtained and the inferences drawn are discussed in Section 4. Section 5 provides the conclusion of this work.

2. Literature Review

The proliferation of hateful content on social media platforms, particularly in the form of multimodal memes, has become a significant societal challenge. The nuanced and contextual nature of memes, which

blend text and imagery, makes automated detection a complex task. This has spurred a considerable body of research focused on developing robust models and comprehensive datasets to identify, understand, and counteract this form of online hate. This survey reviews recent literature, categorizing contributions into three key areas: advancements in multimodal detection architectures, efforts to address data and language specificity, and the extension of Vision-Language Model (VLM) capabilities beyond simple classification.

2.1. Advances in Multimodal Detection Architectures

Early and ongoing research has focused on architecting effective multimodal frameworks that can synergistically analyze both visual and textual components. A foundational approach involves creating hybrid models, such as the Multi-modal Hate Speech Detection Framework (MHSDF), which combines Convolutional Neural Networks (CNNs) for spatial feature extraction with Long Short-Term Memory (LSTM) networks for sequential data analysis across modalities like text, images, and even video. This framework utilizes an attention mechanism to fuse inputs and enhance model interpretability [15]. Other studies have explored pipeline-based systems, for instance by first using Optical Character Recognition (OCR) to extract text, then applying a lexicon-based tool like VADER for sentiment analysis, and finally using a pre-trained CNN to analyze the visual components [16].

More sophisticated deep learning techniques have sought to improve performance by leveraging complex model integrations. One such study proposes a three-stage framework that first generates a textual caption of the meme's image, then processes the multimodal data through an ensemble of three separate transformer-based models to derive a final classification [17]. Another advanced approach utilizes a multi-task learning (MTL) framework, integrating powerful pre-trained models like CLIP, UNITER, and BERT, to concurrently train on four distinct datasets. This method of sharing knowledge across datasets demonstrated state-of-the-art performance over existing unimodal and multimodal techniques [18]. The evolution of this domain is further highlighted by comparative studies evaluating the latest end-to-end Vision-Language Models (VLMs). For example, research fine-tuning the IDEFICS model with a QLoRA strategy has shown its superior accuracy compared to both single-modality models and earlier multimodal methods that rely on separate feature fusion techniques like co-attention [19].

2.2. Specialized Datasets and Methodologies

A significant challenge in hateful meme detection is the scarcity of high-quality, diverse, and context-rich data. Several researchers have focused on creating novel datasets to address these gaps. To tackle data imbalance and move beyond simple binary classification, the Meme-Merge dataset was created to help estimate the severity of offensiveness [20]. Similarly, the GuardHarMem dataset was introduced to provide more nuanced, fine-grained labels for various harm categories like racism and mockery, accompanied by a baseline model, HarMDetect, which integrates auto-generated captions to improve performance [21].

The global nature of social media necessitates models that can function across different languages and cultural contexts, which are often low-resource environments. Research in this area includes the creation of BHM, a novel dataset for Bengali hateful memes annotated not only for hate but also for the specific social entities being targeted, alongside the proposal of the DORA dual co-attention framework [22]. Work has also been done for the Hindi language, involving the creation of a new dataset (with a balanced subset generated via undersampling) and the application of a multimodal Logistic Regression classifier [23]. Similarly, the detection of hateful memes has been introduced as a new research problem for Thai-NLP, with a proposed solution pipeline that links scene text localization, an improved Thai-OCR model, and a multi-task language model trained to handle common misspellings [24]. Beyond language, cultural specificity is also critical, as shown in a study focusing on the Singaporean context, which curated a large-scale dataset labeled by GPT-4V and used it to fine-tune a VLM pipeline for classifying locally nuanced offensive content [25].

Figure 1: Indic Memes (left to right): a) Bengali, b) Bodo, c) Gujarati, d) Hindi

2.3. Extending VLM Capabilities: Explainability, Mitigation, and Critique

With the advent of powerful VLMs, research has begun to explore applications beyond simple detection, focusing on explainability, content mitigation, and critical evaluation of these models' capabilities and safety. One line of work leverages VLMs in a zero-shot setting, using extensive prompt engineering to detect hateful memes without task-specific annotated data, and contributes a typology of common error classes to guide future improvements [26]. To address the "black box" nature of many models, the MemHateCaptioning framework was developed to generate clear, human-like explanations of why a meme is classified as hateful, using a combination of models and Chain-of-Thought (CoT) prompting to improve interpretability [27].

Moving from passive detection to active intervention, the UnHateMeme framework leverages VLMs like GPT-4o to actively mitigate hateful content by replacing toxic visual or textual elements, transforming the meme into a non-hateful version [28]. However, as these models become more capable, their potential for misuse becomes a critical concern. A recent evaluative study of seven different VLMs revealed a significant gap between capability and safety. While the models could often understand the complex cultural and emotional context of hateful memes, they lacked robust safety safeguards, frequently failing to reject hateful prompts and proving vulnerable to misuse for generating new harmful content [29]. This highlights an urgent need for stronger ethical guidelines and safety measures as a key direction for future research.

3. Methodology

The proposed methodology is designed to address the challenges inherent in multi-modal, multi-lingual, and multi-task hate speech classification. Our pipeline is centered around a robust training and evaluation protocol using K-fold cross-validation and a hybrid model architecture that combines a pre-trained vision transformer with one of three powerful, language-specific text encoders.

3.1. Datasets and Pre-Processing

For this study, we utilized four distinct datasets, provided by HASOC-meme 2025, representing four Indic languages: Bangla, Bodo, Gujarati, and Hindi. Example meme images from the four datasets are presented in Figure 1. The datasets provided consisted of train and test subsets consisting of the images and a corresponding csv file. For the train subsets, the csv file contained the labels and OCR text for each image in the subset. For the test subsets, the csv file only contained the OCR text for each image in the subset. The details of each of the datasets are detailed as follows:

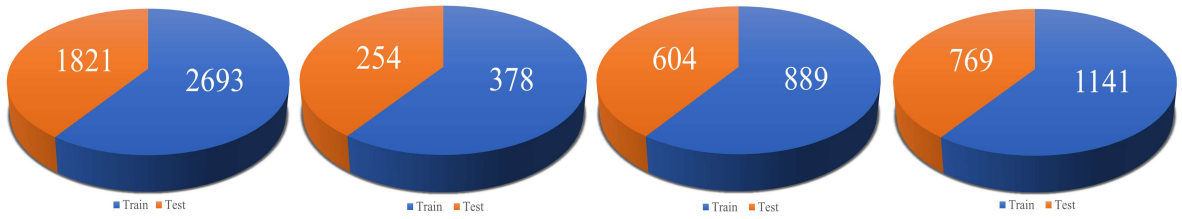

Figure 2: Train-Test split of the Indic datasets (left to right): a) Bangla, b) Bodo, c) Gujarati, d) Hindi

- Bangla - Consisted of total 4514 samples. Consisted of 2693 samples for training and 1821 samples for testing.
- Bodo - Consisted of total 632 samples. Consisted of 378 samples for training and 254 samples for testing.
- Gujarati - Consisted of total 1493 samples. Consisted of 889 samples for training and 604 samples for testing.
- Hindi - Consisted of total 1910 samples. Consisted of 1141 samples for training and 769 samples for testing.

The train-test split for all four languages are represented in Figure 2.

The model is trained for multi-task classification on the dataset to detect sentiment, abuse, vulgarity, and sarcasm in each meme. However, it is observed that there is severe class imbalance across all four tasks in all four datasets that needed to be addressed. Figure 3 details the class imbalance present in each dataset.

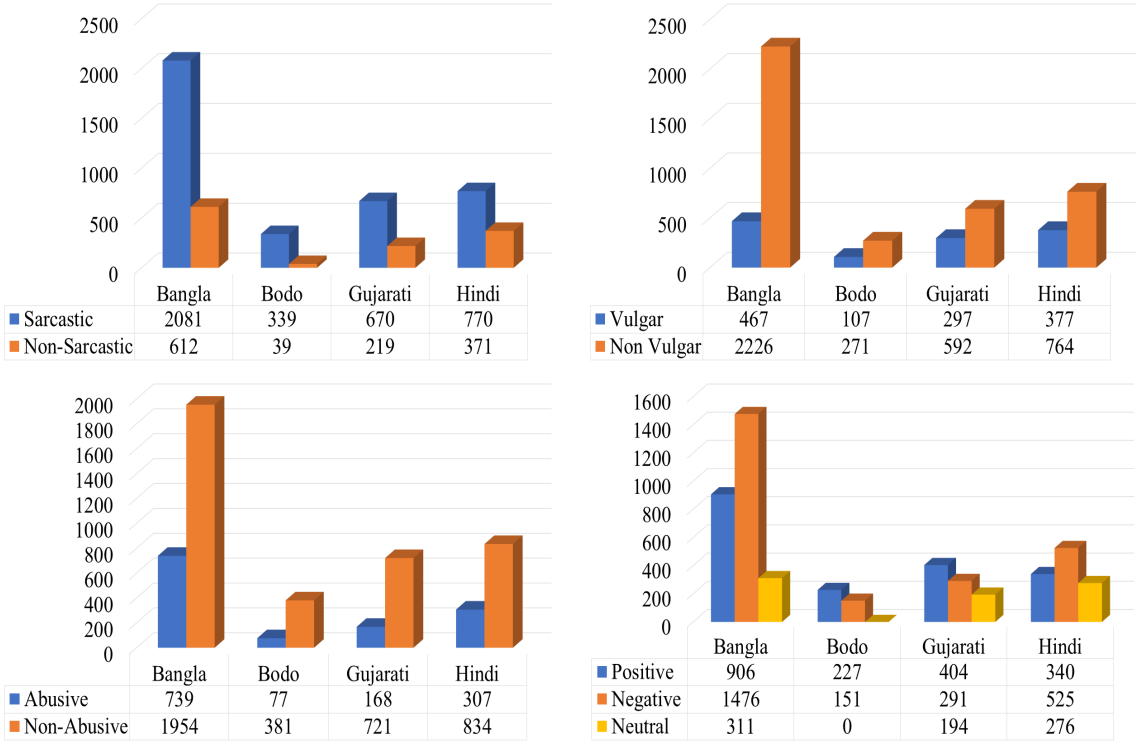

Figure 3: Class Imbalance Graphs (top left to bottom right): a) Bangla b) Bodo c) Gujarati d) Hindi.

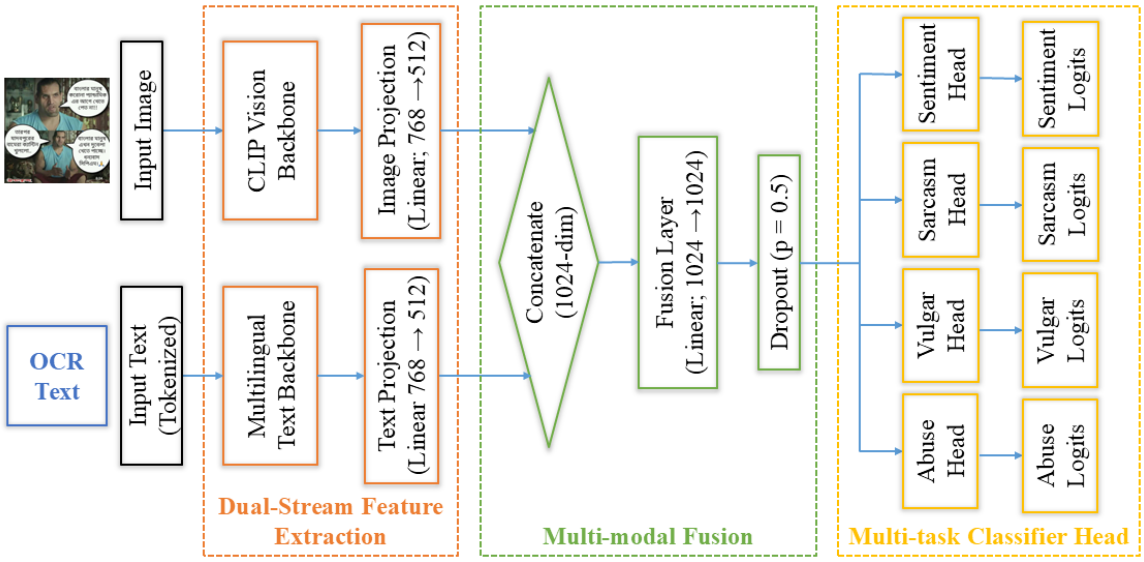

Figure 4: Model Architecture for hate speech and offensive content detection in multi-lingual memes.

3.2. Model Architecture

Our model consists of three primary components: a vision encoder, a text-encoder, and a fusion mechanism that combines their outputs before feeding them to task-specific classification heads. The model architecture is detailed in Figure 4.

3.2.1. Vision Encoder

For the visual modality, the Vision Transformer (ViT-B/32) backbone from the pre-trained CLIP model was utilized. This choice was deliberate and motivated by the unique nature of its training. Unlike traditional vision models pre-trained on fixed-category classification tasks like ImageNet, the CLIP vision encoder was trained via a contrastive objective on 400 million image-text pairs sourced from the internet.

This process compels the model to learn image representations that are deeply aligned with natural language semantics. Consequently, the resulting image features are not merely representations of objects, but also capture the abstract concepts, actions, and sentiments described in the associated text. This semantic richness is particularly advantageous for our task, as the interpretation of memes often depends on the nuanced interplay between visual context and textual content, making CLIP's language-aware features a more powerful starting point than those from standard image classifiers.

The model takes an input image $I \in R^{(H \times W \times C)}$ and produces a 768-dimensional embedding, E_{vision}, from its final layer's pooled output.

3.2.2. Text Encoder

The textual modality, derived from Optical Character Recognition (OCR), presents a unique set of challenges, including noise, non-standard grammar, and the frequent use of code-mixed language (e.g., "Hinglish"). To effectively process the nuances of these Indic languages, our approach is to move beyond generic text models and experiment with and compare three powerful, pre-trained multilingual transformers : MuRIL, XLM-RoBERTa, and mBERT. The goal is to find an encoder that best captures the specific semantic and contextual information required for our classification tasks.

For a given OCR text input, the corresponding tokenizer for the chosen model first converts the text into a sequence of tokens. This sequence is then fed to the text model to generate a 768-dimensional embedding, E_{text}, from its pooled output, which serves as a rich semantic representation of the text.

3.2.3. Multi-Modal Fusion and Classification

The feature vectors from the two encoders are first projected to a harmonized dimension and then fused. The 768-dimensional vision and text embeddings are projected into a 512-dimensional space using separate linear layers:

$$f_{vision} = W_{img} \times E_{vision} + b_{img} \tag{1}$$

$$f_{text} = W_{txt} \times E_{text} + b_{txt} \tag{2}$$

where W_{img} and W_{txt} are the weight matrices and b_{img} and b_{txt} are the biases for their respective linear projection layers. These harmonized 512-dimensional vectors are then concatenated to form a single 1024-dimensional multi-modal feature vector, f_{fused}:

$$f_{fused} = f_{vision} \oplus f_{text} \tag{3}$$

This combined vector is passed through a final fusion block, consisting of a linear layer, a ReLU activation, and a Dropout layer, before being passed to four independent linear classification heads for the final predictions.

3.3. Training and Evaluation

3.3.1. Hold-Out Set and K-Fold Cross Validation

Since for the test set, there were no labels provided, to ensure a robust and unbiased training, and to monitor the models' training, each dataset's training is first split into a training set (90%) and a hold-out test set (10%). We then employ a 5-Fold Stratified Cross-Validation strategy on the 90% training portion. This mitigates the risk of performance variance due to a single, arbitrary data split and allows us to train a robust ensemble of models.

3.3.2. Weighted Loss Function

To address the severe class imbalance observed in the datasets, a Weighted Cross-Entropy Loss is used during training. The weight for each class is calculated as the inverse of its frequency in the training fold. For a given task with C classes, the loss L for a single sample is defined as:

$$L = -w_y log(p_y) \tag{4}$$

where y is the true class index, p_y is the predicted probability for that class, and w_y is the pre-computed weight for class y. This increases the penalty for misclassifying a minority class sample, forcing the model to pay more attention to it.

3.3.3. Ensemble Strategy

The final reported performance is calculated by combining the 5 models trained during cross-validation into an ensemble. For a given sample, the output logits from each of the $K = 5$ models are averaged. The final prediction is the class corresponding to the highest average logit value:

$$logits_{avg} = \frac{1}{K} \sum_{k=1}^{K} logits_k \tag{5}$$

This averaging process improves generalization and produces a more stable final prediction. The final prediction of the ensemble models is made over the hold-out test set.

4. Results and Discussions

The implementation of the proposed methodology pipeline and the results obtained are detailed in this section. The performance of the models are evaluated using the predictions made over the hold-out test set and the actual test set. The performance metrics thus used for evaluation are the Accuracy and Macro F1 Score achieved on the hold-out test set and the Macro F1 Score on the actual test set. The metrics for test set were obtained directly through submission of the test classification file, generated by the models post-training, at the HASOC 2025 portal on the Kaggle platform.

4.1. Implementation

The model's vision and text backbones are initialized using pre-trained weights from CLIP's ViT and the respective language models from Hugging Face, while the custom projection and classification layers are initialized randomly. All input images are resized to a 224×224 dimension by the CLIP processor. For data augmentation during training, several transformations are applied including random horizontal flips, random rotations up to 10 degrees, and color jittering.

The feature vectors from the vision and text encoders are projected and fused into a 1024-dimensional vector. To mitigate overfitting, a dropout rate of 0.5 is applied to this fused vector before it is passed to the final classification heads. The optimization is handled by the Adam optimizer, utilizing a differential learning rate strategy. The pre-trained CLIP vision backbone was fine-tuned with a learning rate of 1×10^{-6}, the language model backbone with 2×10^{-5}, and the custom, randomly initialized layers with 1×10^{-4}. The model was trained for a maximum of 25 epochs with a mini-batch size of 32, using a weighted cross-entropy loss function to address class imbalance.

The entire framework is developed in Python using the PyTorch and Transformers libraries. All experiments are conducted on a standard computing system equipped with an NVIDIA A100 40GB GPU.

4.2. Overall Performance

The proposed strategy is evaluated using quantitative measures with the macro F1 score, achieved on the test set, being the primary performance metric. Due to the datasets being highly imbalanced for all four tasks of Abuse, Sarcasm, Vulgarity, and Sentiment, Macro F1 score is the key performance metric as it evaluates the performance of the models across all classes equally. Table 1 details the results of the experimentation with the CLIP's ViT as the vision encoder and the MuRIL, XLM-RoBERTa, and mBERT as the text encoders respectively. While it is observed that the mBERT pipeline showed the best results on the hold-out test set, the XLM-RoBERTa outperformed the other models on the actual test set achieving the best Macro F1 Score on all four datasets with the highest score being achieved on the Bangla dataset. This proves that the XLM-RoBERTa is the best in learning and generalization of the tasks.

Table 1
Performance Comparison of VLM Models Across Four Languages (in percentage). HOT denotes the Hold-Out Test set.

	HOT Accuracy			HOT Macro F1			Test Macro F1		
Dataset	**mBERT**	**MuRIL**	**XLM-R**	**mBERT**	**MuRIL**	**XLM-R**	**mBERT**	**MuRIL**	**XLM-R**
Bangla	81.65	79.83	82.10	78.40	73.61	77.10	59.56	60.93	**61.82**
Bodo	84.00	87.50	84.00	73.98	74.48	66.33	59.00	57.14	**59.97**
Gujarati	81.53	81.53	80.41	79.02	78.40	77.46	59.32	**60.02**	59.49
Hindi	80.10	82.24	79.61	78.36	80.21	77.57	56.50	56.65	**56.78**

For further delving into the performance of each model pipeline, their performance for each task on the hold-out test set is evaluated. These findings are outlined in Table 2-5.

Table 2
Models' Performance on the Sentiment Task.

	mBERT		MuRIL		XLM-Roberta	
Language	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**
Bangla	71.90	67.40	70.09	66.06	69.79	64.85
Bodo	74.00	70.60	74.00	69.61	76.00	71.43
Gujarati	69.37	69.36	62.16	62.12	66.67	66.80
Hindi	71.71	68.46	71.05	69.99	67.76	66.54

Table 3
Models' Performance on the Sarcasm Task.

	mBERT		MuRIL		XLM-Roberta	
Language	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**
Bangla	88.52	87.90	86.40	85.94	87.31	86.83
Bodo	100.00	100.00	100.00	100.00	100.00	100.00
Gujarati	89.19	89.04	83.78	83.72	90.09	90.01
Hindi	88.82	88.82	87.50	87.50	86.84	86.83

Table 4
Models' Performance on the Vulgar Task.

	mBERT		MuRIL		XLM-Roberta	
Language	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**
Bangla	87.92	75.20	81.87	68.21	88.22	77.22
Bodo	80.00	58.47	78.00	38.25	78.00	38.25
Gujarati	81.98	78.80	81.98	78.80	81.08	77.55
Hindi	80.92	78.50	83.55	80.86	82.24	78.83

Table 5
Models' Performance on the Abuse Task.

	mBERT		MuRIL		XLM-Roberta	
Language	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**
Bangla	84.29	79.10	83.08	78.78	83.99	79.99
Bodo	86.00	64.25	82.00	60.21	84.00	62.12
Gujarati	88.29	80.56	85.59	73.48	87.39	77.77
Hindi	87.50	85.62	86.84	83.30	86.18	83.09

5. Conclusion

In this paper, we conducted a comprehensive study on multi-modal hate speech detection in Indic memes by systematically evaluating dual-encoder architectures combining CLIP with M-BERT, XLM-R, and MuRIL across four languages. Our experiments, grounded in a robust 5-fold cross-validation protocol, establish crucial performance benchmarks. A consistent pattern emerged: while models adeptly classify sentiment, their performance is substantially lower on nuanced, context-dependent tasks like sarcasm, with macro F1 scores on the test set plateauing in the 0.56-0.61 range. This suggests that the features required to discern complex social phenomena are not easily captured by current state-of-the-art models on this type of real-world social media data, thereby highlighting the inherent difficulty of the task.

Potential avenues for future work could therefore shift from model-centric adjustments toward data-centric approaches. Exploring targeted data enrichment, curation, and balancing strategies presents a

promising path for surpassing the current performance ceiling in this challenging domain.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] 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/.

[2] 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.

[3] 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.

[4] 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.

[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, S. Majumdar, A. Chakraborty (Eds.), Forum for Information Retrieval Evaluation (Working Notes) (FIRE 2025) December 17-20, Varanasi , India, CEUR-WS.org, 2025.

[6] Koyel Ghosh and Mithun Das and Sumukh Patel and Nilotpal Bhandary and Alloy Das 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: 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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 (2025). URL:

https://www.sciencedirect.com/science/article/pii/S0167865525003150. doi:https://doi.org/10.1016/j.patrec.2025.09.004.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PmLR, 2021, pp. 8748–8763.

[12] S. Khanuja, D. Bansal, S. Mehtani, S. Khosla, A. Dey, B. Gopalan, D. K. Margam, P. Aggarwal, R. T. Nagipogu, S. Dave, et al., Muril: Multilingual representations for indian languages, arXiv preprint arXiv:2103.10730 (2021).

[13] 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, arXiv preprint arXiv:1911.02116 (2019).

[14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.

[15] R. Prabhu, V. Seethalakshmi, A comprehensive framework for multi-modal hate speech detection in social media using deep learning, Scientific Reports 15 (2025) 13020.

[16] R. Kiruthika, G. Santhosh, R. Santhosh, S. Surya, K. T. Amudhan, Ai-powered system for detecting offensive meme texts on social media, in: 2025 3rd International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA), IEEE, 2025, pp. 1–5.

[17] S. Ivan, T. Ahmed, S. Ahmed, M. H. Kabir, A vision-language multimodal framework for detecting hate speech in memes, in: 2024 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), IEEE, 2024, pp. 464–468.

[18] P. Kapil, A. Ekbal, A transformer based multi task learning approach to multimodal hate speech detection, Natural Language Processing Journal 11 (2025) 100133.

[19] B. Zhang, J. Xu, B. Ihnaini, A. R. Sangi, Beyond text or image: A comparative study of fine-tuned vlm for meme hate speech detection, in: 2025 6th International Conference on Computer Vision, Image and Deep Learning (CVIDL), IEEE, 2025, pp. 414–418.

[20] A. Alzu'bi, L. Bani Younis, A. Abuarqoub, M. Hammoudeh, Multimodal deep learning with discriminant descriptors for offensive memes detection, ACM Journal of Data and Information Quality 15 (2023) 1–16.

[21] S. El-amrany, S. Lamsiyah, M. R. Brust, P. Bouvry, Guardharmem and harmdetect: a multimodal dataset and benchmlark model for fine-grained harmful meme classification, Social Network Analysis and Mining 15 (2025) 63.

[22] E. Hossain, O. Sharif, M. M. Hoque, S. M. Preum, Deciphering hate: identifying hateful memes and their targets, arXiv preprint arXiv:2403.10829 (2024).

[23] K. Dubey, V. Srivastava, G. Sharma, N. Sharma, D. Sharma, U. Ghosh, O. Alfarraj, A. Tolba, Multimodal detection of offensive content in hindi memes, ACM Trans. Asian Low-Resour. Lang. Inf. Process. (2025). URL: https://doi.org/10.1145/3717611. doi:10.1145/3717611, just Accepted.

[24] L. Mookdarsanit, P. Mookdarsanit, Combating the hate speech in thai textual memes, Indonesian Journal of Electrical Engineering and Computer Science 21 (2021) 1493–1502.

[25] C. Yuxuan, W. Jiayang, A. C. L. Chuen, B. S. Guanrong, T. L. C. Jen, S. C. Z. Shen, Detecting offensive memes with social biases in singapore context using multimodal large language models, arXiv preprint arXiv:2502.18101 (2025).

[26] N. Rizwan, P. Bhaskar, M. Das, S. S. Majhi, P. Saha, A. Mukherjee, Exploring the limits of zero shot vision language models for hate meme detection: The vulnerabilities and their interpretations, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 19, 2025, pp. 1669–1689.

[27] R. Sood, A. Anaissi, W. Huang, A. Braytee, Memhatecaptioning: Enhancing hate speech detection in memes with context-aware captioning and chain-of-thought, in: Companion Proceedings of the ACM on Web Conference 2025, 2025, pp. 2034–2041.

[28] M.-H. Van, X. Wu, Detecting and mitigating hateful content in multimodal memes with vision-language models, arXiv preprint arXiv:2505.00150 (2025).

[29] Y. Ma, X. Shen, Y. Qu, N. Yu, M. Backes, S. Zannettou, Y. Zhang, From meme to threat: On the hateful meme understanding and induced hateful content generation in open-source vision language models, in: USENIX Security Symposium (USENIX Security). USENIX, 2025.