Towards Safer Social Media: Multimodal Hate Speech Detection in Memes across Diverse Indian Languages

Rachana Nagaraju[*,†], Hosahalli Lakshmaiah Shashirekha[†]

Department of Computer Science, Mangalore University, Mangalore, Karnataka, India

Abstract

The proliferation of hateful and offensive content on social media raises significant societal concerns, particularly when such messages are conveyed through multimodal memes that combine text and images. Unlike pure textual posts, memes often exploit the interplay between modalities, making the detection of toxic content more challenging. The HASOC-Meme 2025 shared task at FIRE 2025 introduces benchmark datasets in four low-resource languages: Bangla, Hindi, Gujarati, and Bodo, with the objective of identifying hate speech and offensive content by jointly analyzing textual and visual signals embedded in the memes. In this paper, we - team **MUCS** describe our proposed models submitted to HASOC-Meme 2025 shared task. To tackle the challenges, we have developed multimodal models that integrate transformer-based text encoders (*Indic-BERT, MuRIL, XLM-Roberta*) with convolutional and transformer-based vision models (*ResNet, EfficientNet, ViT*) using two fusion mechanisms - concatenation and attention-based strategies, to effectively capture the complementary cues from both modalities. The shared task is formulated as a multi-task learning problem with three binary classification problems of: i) detecting abuse, ii) assessing vulgarity, and iii) accessing sarcasm, and two multi-class classification problems of: i) assigning one of three sentiment labels and ii) identifying one of many targeted communities. This multi-task setup reflects the heterogeneous nature of offensive content in memes: while sentiment span multiple levels of polarity, other categories naturally align with binary distinctions. By jointly optimizing these complementary objectives within a unified architecture, the model is able to leverage shared multimodal representations while also specializing each subtask, thereby improving overall robustness and generalization across languages. Our models achieve competitive performance across languages, ranking 11th in Bangla (macro F1 score 0.5379), 14th in Hindi (macro F1 score 0.5250), 3rd in Gujarati (macro F1 score 0.6185), and 12th in Bodo (macro F1 score 0.5522). These results highlight the effectiveness of multimodal architectures for offensive content identification in memes and demonstrate their adaptability across linguistically diverse and resource-scarce settings.

Keywords

Multimodal learning, Hate speech detection, Offensive content identification, Memes, Low-resource languages

1. Introduction

The exponential growth of user-generated content on social media platforms such as Twitter, Facebook, and Instagram, enables millions of users to express their opinions, share experiences, and engage in public discourse. However, this digital democratization also provides a fertile ground for the dissemination of harmful material, including hate speech, abusive language, and offensive content. Such toxic communication not only marginalizes vulnerable groups but also undermines the quality of online discourse and, by extension, democratic processes [1]. While automated hate speech detection is extensively studied in textual data [2, 3], detecting hateful and offensive memes is still in its infancy. Memes are multimodal artifacts that combine images with overlaid or accompanying text and represent a growing trend on social media. They often rely on the interplay between image and text modalities to convey humor, sarcasm, or offensive undertones. For instance, textual content may appear benign in isolation, but when paired with a culturally or politically loaded image, it can communicate targeted hate. Some sample memes from the HASOC-Meme 2025 dataset are illustrated in Figure 1, highlighting the diversity of languages. This multimodal nature makes hate speech detection substantially more

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

[*]Corresponding author.

[†]These authors contributed equally.

rachananagaraju20@gmail.com (R. Nagaraju); hlsrekha@mangaloreuniversity.ac.in (H. L. Shashirekha)

0000-0002-9421-8566 (H. L. Shashirekha)

(a) Bangla meme (b) Hindi meme (c) Gujarati meme (d) Bodo meme

Figure 1: Sample Memes from HASOC-Meme 2025 Dataset

challenging than unimodal tasks, as it requires models to jointly interpret both textual and visual signals [4, 5].

The HASOC-Meme 2025[1] [6, 7] shared task at FIRE 2025[2] is organized with the goal of benchmarking multimodal approaches for hate speech and offensive content identification in memes in four Indian languages: Bangla, Hindi, Gujarati, and Bodo. These languages are underrepresented due to a lack of linguistic resources and computational tools, making the shared task particularly important for advancing research in low-resource settings. The shared task involves analyzing multimodal data (image and text) to detect abuse, identify targeted communities, assess vulgarity and sarcasm, and assign sentiment labels, in the given meme.

In this paper, we - team **MUCS** describe the models submitted to HASOC-Meme 2025 shared task. To address the challenges of multimodal hate speech detection, we designed models that leverage state-of-the-art transformer-based text encoders—Indic-BERT [8], MuRIL [9], and XLM-RoBERTa [10]—along with deep vision backbones—ResNet [11], EfficientNet [12], and Vision Transformers (ViT) [13], to represent text and image respectively. The shared task is formulated as a multi-task learning problem with three binary classification problems of: i) detecting abuse, ii) assessing vulgarity, and iii) accessing sarcasm, and two multi-class classification problems of: i) assigning one of three sentiment labels and ii) identifying one of many targeted communities. We explore two fusion strategies: concatenation-based and attention-based fusion, to integrate features from image and text modalities effectively. The proposed models deliver competitive performance, ranking 11^{th} in Bangla with a macro F1 score of 0.5379, 14^{th} in Hindi with 0.5250, 3^{rd} in Gujarati with 0.6185, and 12^{th} in Bodo with 0.5522. These results demonstrate both the promise and challenges of multimodal learning for hate speech meme detection in linguistically diverse and resource-constrained environments. By contributing a systematic exploration of multimodal architectures and fusion strategies, this work advances the development of robust content moderation systems for low-resource languages.

The subsequent sections of this paper details the related works (Section 2), methodology (Section 3), experiments, results, and implications of our approach (Section 4) followed by conclusion and future works (Section 5).

2. Related Works

The growing presence of multimodal hateful and offensive content in Indian languages poses unique challenges compared to English and other high-resource languages. While earlier works demonstrate the promise of multimodal fusion techniques, the complexity of code-switching, script diversity, and cultural nuances in memes make this task particularly demanding.

Dubey et al. [14] focus on detecting offensive content in Hindi memes, highlighting the need for automated solutions in low-resource languages. Their approach combines textual and visual cues

[1] https://hasocfire.github.io/hasoc/2025/
[2] https://fire.irsi.org.in/fire/2025/home

through Logistic Regression classifier, achieving an accuracy of 81%. Although their work demonstrates that multimodal fusion can significantly outperform unimodal methods in Hindi, the reliance on shallow classifiers restricts scalability and generalization compared to transformer-based approaches. Karim et al. [15] address the challenge of hate speech detection in Bengali by systematically exploring multimodal architectures. They use recurrent neural networks and pretrained language models such as BanglaBERT and XLM-RoBERTa, alongside deep visual encoders including ResNet-152 and DenseNet-161. Their best multimodal fusion model reaches an F1 score of 0.83, outperforming text-only or image-only baselines. The study highlights the complementary value of vision and text, though fusion gains are relatively modest, pointing to the difficulty of balancing modalities. Hossain et al. [16] investigate meme classification in Bengali and code-mixed contexts, by integrating CNN-based visual models with transformer-based text encoders. They show that multimodal pipelines improve F1 scores by about 3% compared to unimodal setups. Their work emphasizes that visual cues often provide disambiguating context when textual content alone is insufficient, though the improvements come with the expense of computational complexity.

Debnath et al. [17] extend the scope of hateful meme detection by examining advanced multimodal architectures to handle context-rich Bengali memes. They evaluate BanglaBERT with visual backbones such as ResNet, Inception, and Vision Transformer, and found that multimodal models consistently outperform unimodal ones with accuracies around 64%. Despite these improvements, the study underscores challenges in aligning visual and textual features, which often limits overall performance gains. Rajput et al. [18] concentrate on politically motivated and code-switched Indian memes, a domain that presents unique linguistic and cultural challenges. They develop a CNN + LSTM hybrid architecture, where CNNs extract visual features and LSTMs model text sequences. Their approach achieves state-of-the-art results on their benchmark dataset, demonstrating the strength of hybrid neural designs. However, the system is less adaptable to broader meme domains outside political discourse.

Manukonda and Kodali [19] examine misogyny detection in Tamil and Malayalam memes, emphasizing the under representation of Dravidian languages in multimodal hate speech research. They propose a transliteration-aware XLM-RoBERTa encoder for text, fused with ResNet-50 image embeddings through an attention-BiLSTM module. The system delivers strong results with macro-F1 scores of 0.8805 for Malayalam and 0.8081 for Tamil. Despite its effectiveness, the study points out difficulties with class imbalance and limited generalization across visual domains. Wong and Durward [20] explored target offensive content detection in Gujarati and Hindi as part of the LT-EDI-2024[3] shared task. Their system leveraged transformer-based classifiers with explicit handling of code-mixing and script-switching, ranking second in Gujarati and Telugu subtasks. The work confirms transformers as effective baselines for under-resourced Indian languages, though noise from OCR and inconsistent transliteration practices remain limiting factors.

Overall, existing studies on hate and offensive meme detection in Indian languages clearly establish the importance of multimodal approaches for hateful meme detection and also reveal persistent challenges in modality alignment, code-mixing, and domain adaptation. These insights directly motivate our work, where we design and evaluate robust multimodal fusion strategies to advance hate speech and offensive content detection in Bangla, Hindi, Gujarati, and Bodo memes. By leveraging recent transformer-based language models and modern visual encoders, combined with popular fusion strategies, we aim to contribute to more robust hate speech meme detection in multilingual social media contexts.

3. Methodology

Hate speech meme detection is formulated as a multi-task problem, with three binary classification problems of: i) detecting abuse, ii) assessing vulgarity, and iii) accessing sarcasm, and two multi-class classification problems of: i) assigning one of three sentiment labels and ii) identifying one of many targeted communities, leveraging text (OCRed content or caption) and image pair as inputs. The overall workflow consists of text pre-processing, image pre-processing, feature representation and

[3]https://sites.google.com/view/lt-edi-2024

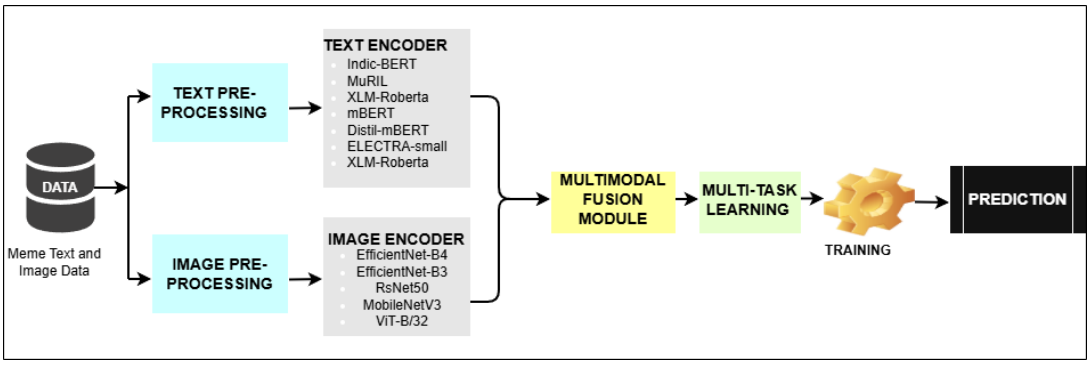


Figure 2: Overview of the Proposed Multi-task Learning for HASOC-Meme Detection

fusion, hyperparameter configuration, multi-task formulation, and evaluation. The proposed multi-task learning for HASOC-Meme detection is illustrated in Figure 2 and details of the steps involved are given below:

3.1. Text Pre-processing

The following steps are applied to prepare textual inputs:

- **Text normalization:** text data is normalized to lowercase and `id`/`image_id` are unified into `ids`.
- **Schema validation:** an exception is raised if no OCR/text field exists, ensuring consistency across datasets.
- **Tokenization:** tokenizers from text pretrained models are applied with a maximum sequence length of 128 tokens, and all input sequences are padded to this fixed length.

3.2. Image Pre-processing

The following steps are applied to prepare image inputs:

- **Decoding:** images are loaded in RGB mode using the `PIL` library[4].
- **Normalization:** all images are resized to 224×224 pixels and pixel values are normalized with a mean and standard deviation of 0.5 for each channel.
- **Augmentation:** during training, random horizontal flips ($p = 0.5$) and random brightness/contrast adjustments ($p = 0.2$) are applied.
- **Path resolution:** image identifiers are matched with both ".jpg" and ".png" extensions; otherwise, a zero tensor of size $3 \times 224 \times 224$ is substituted to maintain batch consistency.

3.3. Feature Representation and Fusion

Text and image feature representations are carried out independently for text and images, respectively, followed by their fusion to obtain joint representations for image and text pairs as given below:

- **Text feature representation:**
  - Text embeddings are obtained from the following pretrained transformer-based encoders:

[4]https://pillow.readthedocs.io/en/stable/reference/index.html

    - **Indic-BERT** - is a multilingual transformer model pre-trained on 12 major Indian languages and English using a Masked Language Modeling (MLM) objective. It captures rich linguistic and semantic features across related Indic languages, making it suitable for cross-lingual and multilingual NLP tasks in the Indian context.
    - **Multilingual BERT (mBERT)** - is a multilingual version of BERT trained on Wikipedia text from 104 languages. It uses a shared WordPiece vocabulary and a single transformer network to model multiple languages simultaneously and demonstrates impressive zero-shot cross-lingual transfer capabilities, enabling it to perform well on languages it was not fine-tuned on.
    - **Multilingual Representations for Indian Languages (MuRIL)** - is developed by Google to improve multilingual understanding of Indian languages. Unlike mBERT, MuRIL is trained on both monolingual and transliterated text, as well as parallel corpora. This helps it to better capture code-mixing and translation nuances common in Indian language text.
    - **Distil-mBERT** - is a distilled version of mBERT that retains 95% of mBERT's performance while being 40% smaller and 60% faster. It is trained using knowledge distillation, where the smaller model learns to mimic the behavior of mBERT, making it efficient for resource-constrained environments.
    - **ELECTRA-small** - is a lightweight transformer-based language model trained using the Replaced Token Detection (RTD) objective, which makes it more sample-efficient than traditional MLM models. Instead of masking and predicting tokens, ELECTRA trains a discriminator to detect replaced words, resulting in better performance with fewer computational resources.
    - **XLM-Roberta** - is a multilingual variant of RoBERTa trained on 2.5 TB of filtered CommonCrawl data across 100 languages. It improves upon mBERT by leveraging more data, longer training, and dynamic masking, achieving state-of-the-art results on multiple cross-lingual benchmarks.
  - The [CLS] token embedding (`last_hidden_state[:,0,:]`) is extracted as the global sentence-level representation.
  - A linear adapter layer projects the text embedding into a 512-dimensional space, followed by ReLU activation and dropout ($p = 0.1$).

- **Image feature representation:**
  - Visual embeddings are obtained from the following pretrained encoders:
    - **ResNet50** - is a 50-layer deep Convolutional Neural Network (CNN) that introduced residual learning through skip connections. These residual blocks help train very deep networks efficiently by addressing the vanishing gradient problem, making ResNet50 a standard backbone for many computer vision tasks.
    - **EfficientNet-B3** - belongs to the EfficientNet family which scales network depth, width, and resolution using a compound coefficient. It achieves high accuracy with optimized computational efficiency, outperforming many larger models with significantly fewer parameters.
    - **EfficientNet-B4** - is a deeper and wider version of EfficientNet-B3, offering improved representational power while maintaining strong efficiency. It balances accuracy and computational cost, making it suitable for applications requiring higher performance with moderate resource constraints.
    - **MobileNetV3** - is an efficient CNN architecture optimized for mobile and edge devices. It combines depthwise separable convolutions, squeeze-and-excitation modules, and a

lightweight neural architecture search design and achieves a strong trade-off between speed and accuracy.

* **Vision Transformer (ViT-B/32)** - is a transformer-based architecture for image understanding. It divides an image into fixed-size patches, linearly embeds them, and processes them using standard transformer layers. ViT-B/32 uses a patch size of 32×32 and achieves competitive results compared to convolutional models, especially when trained on large datasets.

- Features are extracted from the penultimate layer by setting `num_classes=0`, which returns globally pooled features.
- The resulting feature vector dimensionality depends on the backbone.
- A linear adapter layer projects the image embedding into a 512-dimensional space, followed by `ReLU` activation and dropout ($p = 0.1$).

- **Fusion strategies:**
  - **Concatenation:** text and image embeddings are concatenated into a single vector.
  - **Attention:** text and image embeddings are projected into a shared 512-dimensional space. Multi-head attention (4–8 heads) is applied in both directions, enhancing text with image context and image with text context. The resulting enhanced embeddings are concatenated.
  - **Text and Vision Encoder Combinations:** the following text–image encoder combinations are explored:
    1. `Indic-BERT` + `EfficientNet-B4`
    2. `MuRIL` + `EfficientNet-B3`
    3. `XLM-Roberta` + `ResNet50`
    4. `mBERT` + `ResNet50`
    5. `Distil-mBERT` + `EfficientNet-B3`
    6. `ELECTRA-small` + `MobileNetV3`
    7. `XLM-Roberta` + `ViT-B/32`

The fused image and text representation is used to train a multi-task model for HASOC-Meme detection task.

3.4. Task Formulation

The HASOC-Meme detection problem is formulated as a multi-task learning problem with three binary classification problems and two multi-class classification problems, and the formulation of the task is described below:

- **Multi-class Classification**

i) **Sentiment Analysis** - has three categories: {Negative, Neural, and Positive}, mapped to numeric encodings {`0`, `1`, and `2`}, respectively. The predictions for the test data are generated using a sigmoid activation over one output logit, where a value above 0.5 indicates a positive sentiment.

ii) **Target Community Identification** - has many class labels and the number of labels varies from one language to another and these labels are encoded numerically. Some of the labels common to all the languages are given below:

- Gender - Any reference to male, female, non-binary, or transgender identities.
- Religion - Mentions or imagery related to any religious belief, deity, or practice.

Table 1
Hyperparameters and architectural settings

Item	Setting
Text max length	128 tokens (padding + truncation)
Epochs	3
Optimizer	AdamW
Learning rate	1×10^{-5} or 2×10^{-5}
Validation split	20% (15% with stratification)
Image size	224×224
Normalization	mean = std = 0.5 per channel
Augmentations	flip ($p = 0.5$), brightness/contrast ($p = 0.2$)
Text encoders	Indic-BERT, mBERT, MURIL, Distil-mBERT, ELECTRA-small, XLM-Roberta
Image encoders	ResNet50, EfficientNet-B3/B4, MobileNetV3, ViT-B/32
Fusion	concatenation or attention (4–8 heads)
Adapters	Linear $\rightarrow$ 512 with ReLU + Dropout (0.1)
Classifier (sentiment)	Linear (256$\rightarrow$128) $\rightarrow$ ReLU $\rightarrow$ Dropout $\rightarrow$ Linear (128$\rightarrow$3)
Classifier (binary)	Linear (256$\rightarrow$128) $\rightarrow$ ReLU $\rightarrow$ Dropout $\rightarrow$ Linear (128$\rightarrow$3)
Loss functions	CrossEntropy (sentiment), BCEWithLogits (binary tasks)
Thresholds	Binary predictions are thresholded at 0.5
Device	CUDA if available; otherwise, CPU

  - Individual - Specifically mentions or portrays a particular person.
  - Political - Targets political ideologies, parties, politicians, or policies.
  - National Origin - Targets people based on their country or ethnicity.
  - Social Sub-groups - Groups based on socio-economic status, occupation, cultural identity, or other affiliations.
  - Others - Any target that does not fall into the above categories.
  - None - If the meme does not target any specific community, no target label is assigned.

  Predictions for the test data are generated using a sigmoid activation over one output logit, where a value above 0.5 indicates a positive sentiment.

- **Binary Classification**
  i) **Sarcasm vs. Non-Sarcasm**
  ii) **Vulgar vs. Non-Vulgar**
  iii) **Abusive vs. Non-Abusive**

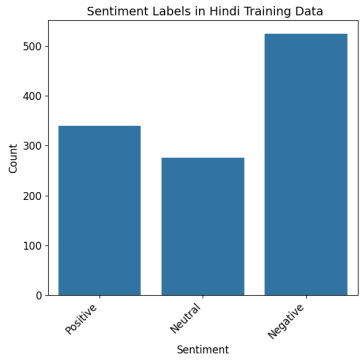
  Labels are encoded as 1 if the corresponding field contains indicators such as `Sarcastic`, `Vulgar`, or `Abusive`, and 0 otherwise. Predictions for test data are generated through sigmoid activations over three independent logits, one for each binary task.

All models are trained under identical conditions to ensure a fair comparison across architectures and fusion strategies. The models are trained with a consistent set of hyperparameters for all experiments and configuration of the hyperparameters is shown in Table 1.

4. Experiments and Results

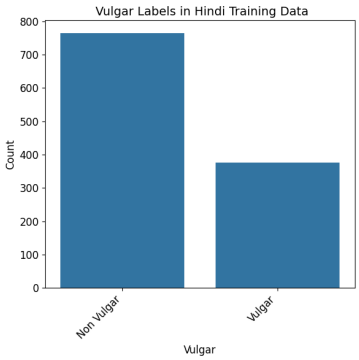
All experiments are done in `PyTorch` with `HuggingFace Transformers`[5] for text encoders and `timm`[6] for image encoders. The size of Train and Test sets for each language in HASOC-Meme 2025

[5]https://huggingface.co/docs/transformers/index
[6]https://huggingface.co/timm

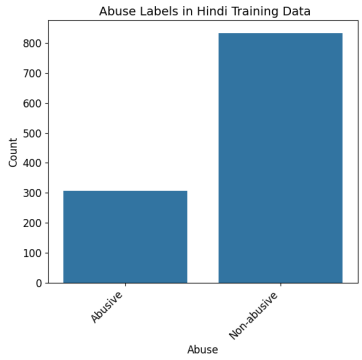


(a) Sentiment Classification

(b) Sarcasm vs. Non-Sarcasm



(c) Vulgar vs. Non-Vulgar



(d) Abusive vs. Non-Abusive

(e) Target Labels

Figure 3: Label-wise Data Distribution in Hindi Dataset

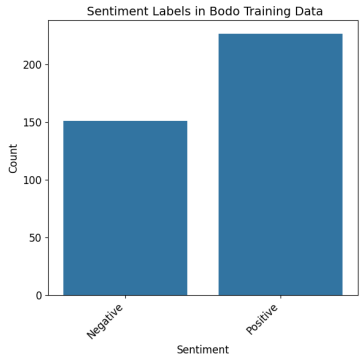
shared task is shown in Table 2 and the distribution of labels for each subtask in Hindi, Bodo, Gujarati, and Bangla are shown in figures 3, 4, 5 and 6, respectively. The text data in these datasets are in native as well as roman script.

Table 2
Size of Train and Test Sets in HASOC-Meme Datasets

Language	Train Set Size	Test Set Size
Bangla	2,693	1,821
Bodo	378	254
Gujarati	889	604
Hindi	1,141	769

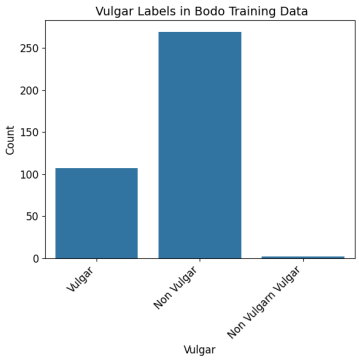
The best performances of the proposed models for the four languages in terms of macro F1 scores are shown in Table 3 and comparison of the performances of proposed models with the other participants' models in HASOC-Meme 2025 shared task for all the four languages is shown in Figure 7. The results indicate that the model performs best on the Gujarati dataset, achieving a macro F1 score of 0.61848 with a rank of 3. This can be attributed to the relatively balanced training data for Gujarati, which probably enables better learning of both textual and visual features. Performances on Bangla, Hindi, and Bodo datasets, are moderate with ranks in the range 11 to 14 and macro F1 scores between 0.52497 and 0.55215. The slightly lower performance on these languages may be due to smaller training datasets, missing annotations, or linguistic and script complexities that pose challenges for text encoding. Overall, these results demonstrate that multimodal approaches can effectively handle meme classification across multiple languages, although dataset size, completeness, and linguistic diversity continue to impact performance.

Beyond the observed class imbalance, several dataset-specific and multimodal challenges contribute

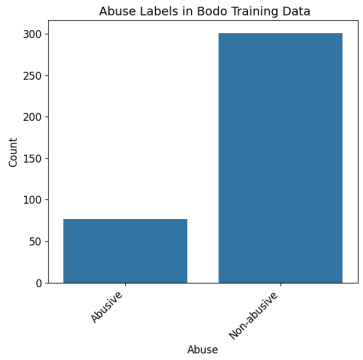


(a) Sentiment Classification

(b) Sarcasm vs. Non-Sarcasm



(c) Vulgar vs. Non-Vulgar



(d) Abusive vs. Non-Abusive

(e) Target Labels

Figure 4: Label-wise Data Distribution in Bodo Dataset

to variations in model performance across languages. In Bangla dataset, the distribution of *Target* class labels exhibit exceptionally high cardinality which influences the learning dynamics. The presence of such intricate label patterns makes representation learning more difficult, forcing the model to cope up with imbalanced and overlapping semantic cues. Consequently, the model struggles to maintain consistent feature alignment across modalities, which reduces its overall macro F1 performance.

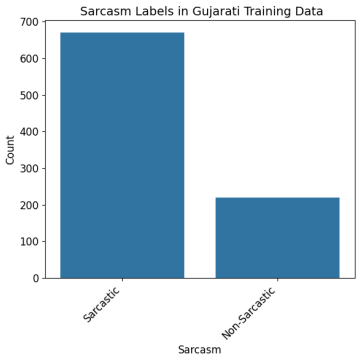
Bodo dataset faces a unique issue with the presence of the "Non-Vulgar and Vulgar" class in the *Vulgar* category. This ambiguous labeling creates inconsistency during training, confusing the model's decision boundaries between clearly defined vulgar and non-vulgar content. The presence of such mixed or mislabeled categories introduces noise, which degrades both convergence stability and classification accuracy in the *Vulgar* task.

A further source of difficulty arises due to the presence of numerous NaN values in the *Target* classification task across all languages. Depending on way such entries are handled—dropped, retained, or ignored—they can alter sample distribution, thereby influencing model generalization and stability during training.

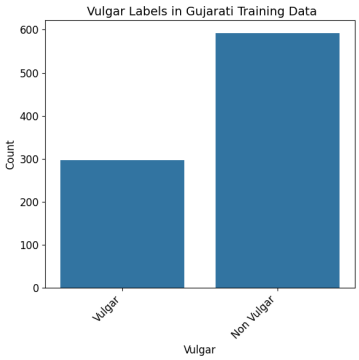
Another contributing factor is the potential mismatch between textual and visual representations across languages due to differences in model architectures and fusion mechanisms. For example, *IndicBERT* and *EfficientNet_B4* with attention-based fusion are used for Hindi, whereas *XLM-RoBERTa* and *ViT-B/32* with concatenation are employed for Gujarati. These combinations may not align multimodal features equally well for all languages, leading to varying degrees of representational synergy. The attention mechanism tends to underperform when visual cues are sparse or semantically weak, while concatenation fusion provides a more robust shared embedding space in certain cases, such as Gujarati.

Sarcasm detection remains a cross-lingual challenge in this study. Despite moderate balance in the *Sarcasm* category, all languages exhibit difficulty in accurately identifying sarcastic expressions. This stems from the subtle and context-dependent nature of sarcasm, which relies heavily on linguistic

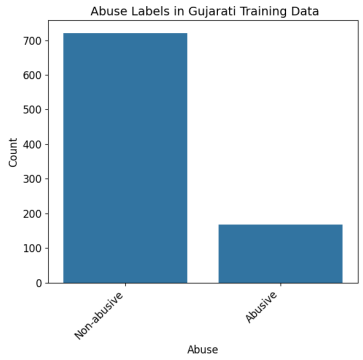
(a) Sentiment Classification



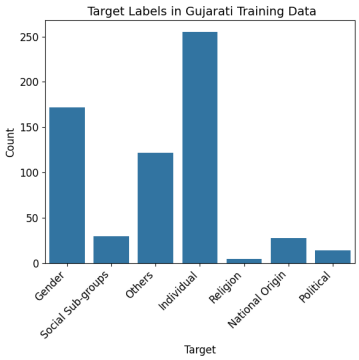
(b) Sarcasm vs. Non-Sarcasm



(c) Vulgar vs. Non-Vulgar



(d) Abusive vs. Non-Abusive



(e) Target Labels

Figure 5: Label-wise Data Distribution in Gujarati Dataset

Table 3
Performance of the proposed models

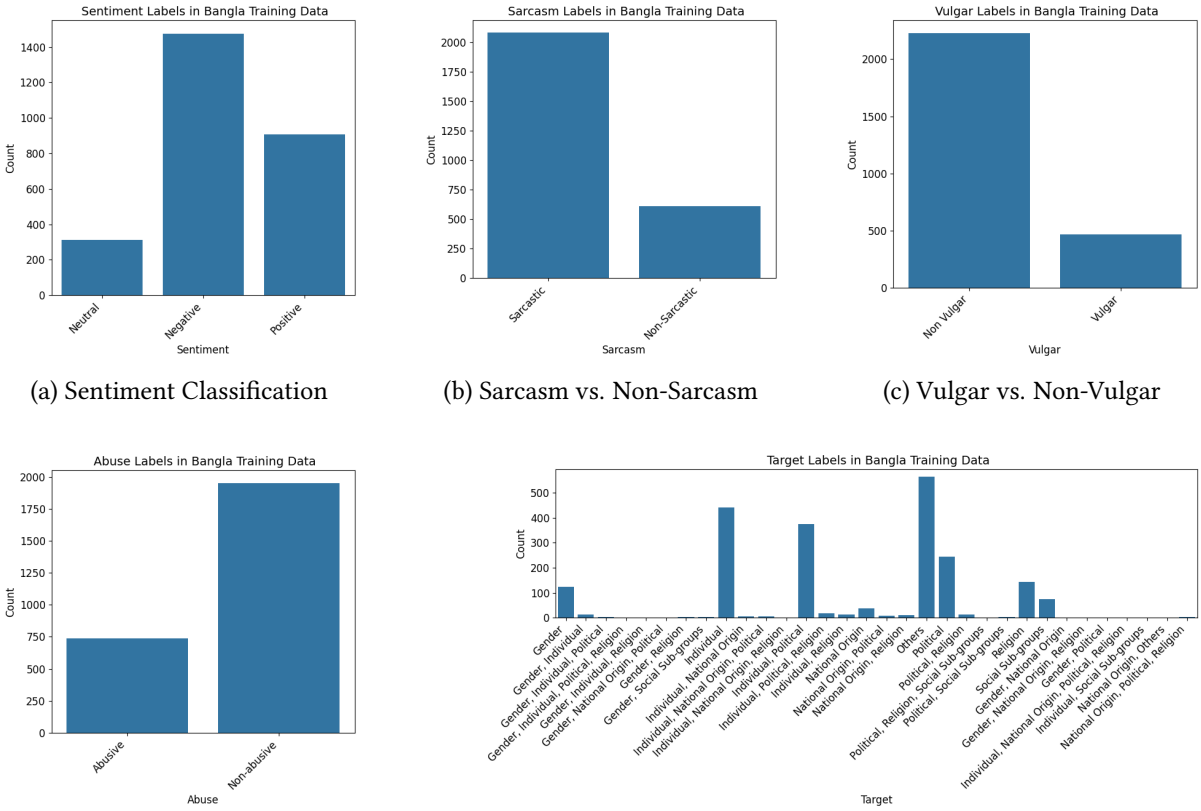
Language	Rank	Macro F1	Text Model	Vision Model	Fusion Type
Bangla	11	0.53785	DistilBERT	EfficientNet_B3	Concatenation
Hindi	14	0.52497	Indic-BERT	EfficientNet_B4	Attention
Gujarati	3	0.61848	XLM-Roberta	ViT-B/32	Concatenation
Bodo	12	0.55215	XLM-Roberta	ViT-B/32	Concatenation

cues and cultural context rather than explicit visual indicators. The visual modality provides limited information for recognizing sarcasm, making the performance heavily dependent on the model's ability to interpret implicit meaning and irony.

Overall, these analyses reveal that performance disparities are not solely driven by class imbalance. They also arise from deeper issues such as complex label structures and missing entries in *Target*, ambiguous or inconsistent annotations in *Vulgar*, suboptimal alignment between text and image encoders, and the inherent linguistic difficulty of sarcasm detection. Addressing these factors through refined data pre-processing, improved annotation quality, and adaptive multimodal fusion strategies can lead to more robust and equitable multilingual meme classification performance.

5. Conclusion and Future Work

In this study, we - team MUCS presented a comprehensive multimodal approach for meme classification across four Indic languages: Bangla, Bodo, Gujarati, and Hindi, as part of the HASOC-Meme 2025 shared task at FIRE 2025. We formulated meme classification as a multi-task learning problem with two multi-class classification objectives and three binary classification objectives. We further explored



(a) Sentiment Classification (b) Sarcasm vs. Non-Sarcasm (c) Vulgar vs. Non-Vulgar

(d) Abusive vs. Non-Abusive (e) Target Labels

Figure 6: Label-wise Data Distribution in Bangla Dataset

two fusion strategies — concatenation and attention-based fusion, to integrate the features from both modalities. Our system effectively leveraged both textual and visual modalities to capture the nuanced semantics of memes. The proposed system achieved competitive results: Rank 3 for Gujarati with a macro F1-score of 0.61848, Rank 11 for Bangla with a macro F1-score of 0.53785, Rank 12 for Bodo with a macro F1-score of 0.55215, and Rank 14 for Hindi with a macro F1-score of 0.52497. These results highlight the model's strong adaptability and effectiveness across diverse linguistic and visual contexts, even in low-resource settings. For future work, we aim to enhance performance through more sophisticated multimodal fusion strategies and domain-adaptive pretraining. Expanding the training data with larger multilingual and cross-domain meme datasets, as well as incorporating contrastive learning and prompt-based fine-tuning techniques, may further improve generalization. Additionally, we plan to explore multimodal transformers explicitly optimized for low-resource Indic languages to achieve deeper semantic alignment between text and image modalities.

Declaration on Generative AI

We acknowledge the use of generative AI tools in supporting certain aspects of this work, such as drafting text, formatting code snippets, and organizing content. However, all experimental design, data processing, model training, and result analysis are conducted by the team. Generative AI is used solely as an assistive tool, and all scientific conclusions, interpretations, and discussions presented in this report are our own.

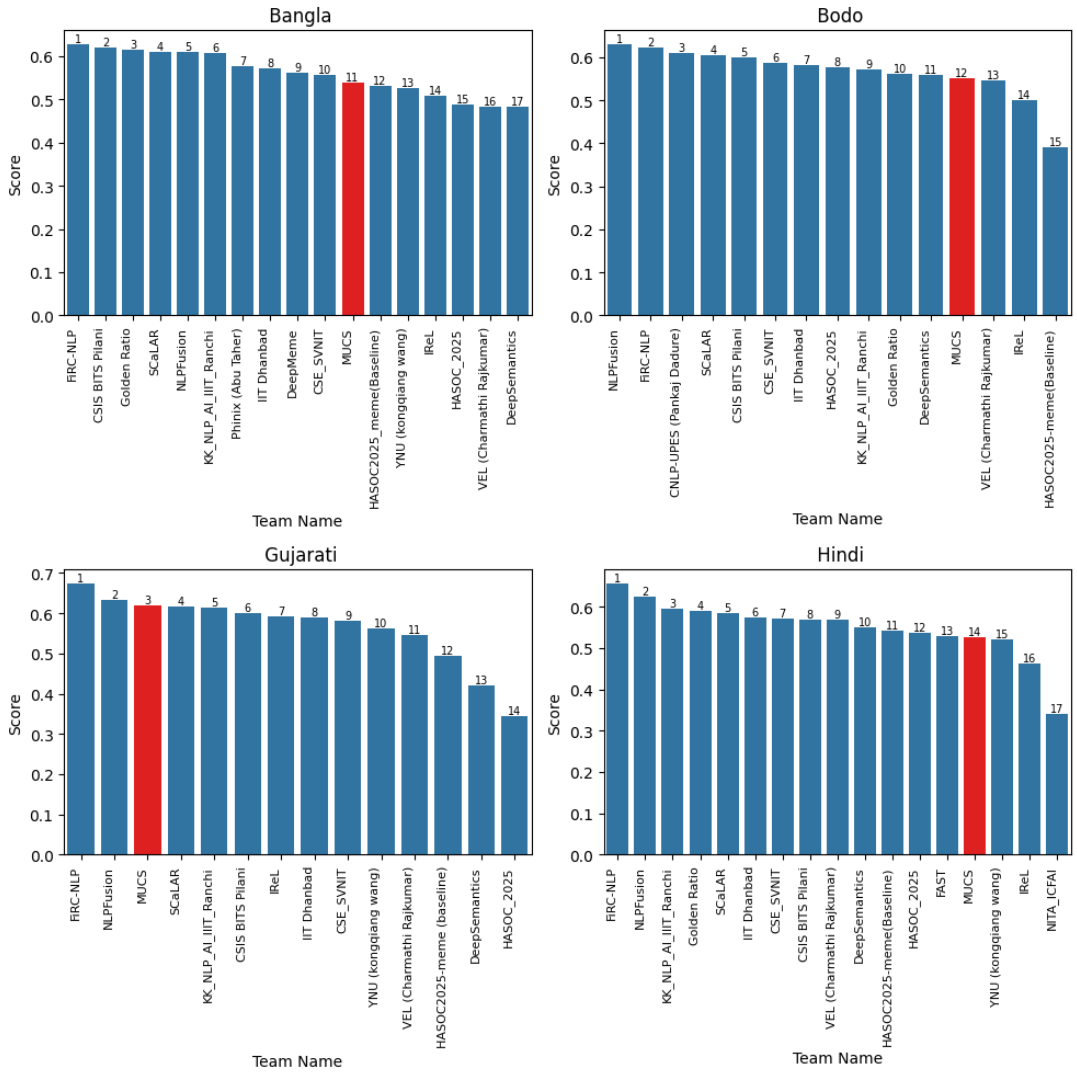


Figure 7: Comparison of the Performances of Proposed Models with the Other Participants' Models in HASOC-Meme Shared Task for all the Four Languages

References

[1] J. Habermas, The Theory of Communicative Action: Reason and the Rationalization of Society, volume 1, Beacon Press, 1984.

[2] A. Schmidt, M. Wiegand, A Survey on Hate Speech Detection Using Natural Language Processing, in: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, 2017, pp. 1–10. doi:10.18653/v1/W17-1101.

[3] P. Fortuna, S. Nunes, A Survey on Automatic Detection of Hate Speech in Text, ACM Computing Surveys 51 (2018) 1–30. doi:10.1145/3232676.

[4] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 9448–9459. URL: https://arxiv.org/abs/2005.04790.

[5] S. Pramanick, S. Sharma, D. Dimitrov, M. S. Akhtar, P. Nakov, T. Chakraborty, MOMENTA: A Multimodal Framework for Detecting Harmful Memes and Their Targets, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 4439–4455. doi:10.18653/v1/2021.findings-emnlp.379.

[6] K. Ghosh, M. Das, M. Narzary, S. Saha, S. Barman, 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: 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, p. N/A.

[7] 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, p. N/A.

[8] D. Kakwani, A. Kunchukuttan, S. M. Golla, et al., IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages, in: Proceedings of the 12th Language Resources and Evaluation Conference (LREC), 2020, pp. 4940–4951. URL: https://aclanthology.org/2020.lrec-1.609.

[9] S. Khanuja, S. Dandapat, A. Srinivasan, et al., MuRIL: Multilingual Representations for Indian Languages, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 2148–2161. doi:10.18653/v1/2021.findings-acl.189.

[10] A. Conneau, K. Khandelwal, N. Goyal, et al., Unsupervised Cross-lingual Representation Learning at Scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451. doi:10.18653/v1/2020.acl-main.747.

[11] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[12] M. Tan, Q. V. Le, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, in: Proceedings of the 36th International Conference on Machine Learning (ICML), 2019, pp. 6105–6114. URL: http://proceedings.mlr.press/v97/tan19a.html.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in: International Conference on Learning Representations (ICLR), 2021, p. N/A. URL: https://arxiv.org/abs/2010.11929.

[14] 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 Transactions on Asian and Low-Resource Language Information Processing (2025). doi:10.1145/3717611, dataset of 9,262 Hindi memes; logistic regression multimodal model with 0.81 accuracy.

[15] M. R. Karim, S. K. Dey, T. Islam, B. R. Chakravarthi, Multimodal Hate Speech Detection from Bengali Memes and Texts, 2022. URL: https://arxiv.org/abs/2204.10196, bengali multimodal dataset; best fusion XLM-R + DenseNet-161 F1 = 0.83.

[16] E. Hossain, O. Sharif, M. M. Hoque, MUTE: A Multimodal Dataset for Detecting Hateful Memes, in: AACL-IJCNLP Student Research Workshop, 2022, p. N/A. 4,158 Bengali and code-mixed memes; multimodal improves 3%.

[17] R. S. Debnath, N. B. Firuj, A. W. Shakib, S. Sultana, M. S. Islam, ExMute: A Context-Enriched Multimodal Dataset for Hateful Memes, in: First Workshop on NLP for Indo-Aryan and Dravidian Languages (IndoNLP), 2025, p. N/A. Context-enriched Bengali hateful meme dataset; multimodal 64% accuracy.

[18] K. Rajput, R. Kapoor, K. Rai, P. Kaur, Hate Me Not: Detecting Hate Inducing Memes in Code Switched Languages, 2022. URL: https://arxiv.org/abs/2204.11356, iPM dataset of Indian political memes; CNN + LSTM model.

[19] D. P. Manukonda, R. G. Kodali, Multimodal Misogyny Meme Detection in Low-Resource Dravidian Languages Using Transliteration-Aware XLM-RoBERTa, ResNet-50, and Attention-BiLSTM, in: DravidianLangTech at NAACL 2025, 2025, p. N/A. Macro-F: 0.8805 (Malayalam), 0.8081 (Tamil).

[20] S. G.-J. Wong, M. Durward, cantnlp@LT-EDI-2024: Automatic Detection of Anti-LGBTQ+ Hate Speech in Under-resourced Languages, arXiv preprint (2024). URL: https://arxiv.org/abs/2401.15777, transformer system; ranked second in Gujarati and Telugu.