Multimodal Hate Speech and Offensive Content Classification in Code-Mixed Indian Language Memes

Sai Saketh Nandam[1], Anand Kumar Madasamy[1]

[1]Department of Information Techonology, National Institute of Techonology Karnataka, Surathakal, P. O. Srinivasnagar, Mangalore - 575 025

Abstract

This paper presents the participation of SCaLAR Team in the HASOC Meme 2025 shared task, which focused on detecting hate speech and offensive content in memes across four Indian languages: Bangla, Hindi, Gujarati, and Bodo. The task consisted of four subtasks—sentiment, sarcasm, vulgarity, and abuse detection—requiring multimodal analysis that integrates both visual and textual cues from images and OCR-extracted text. We experimented with two architectures: a CLIP-based model combining CLIP ViT visual features with multilingual Sentence Transformer embeddings, and a ViT + XLM-Roberta model where visual and textual features were concatenated for classification. The CLIP-based model achieved superior performance in Bangla and Gujarati, underscoring the effectiveness of contrastive pretraining for aligning visual and linguistic representations. Our system ranked 4th in Bangla (F1 = 0.6103), 5th in Hindi (F1 = 0.5847), 4th in Gujarati (F1 = 0.6172), and 4th in Bodo (F1 = 0.6039), consistently placing in the top five across all languages. These results demonstrate the robustness of multimodal learning approaches for meme classification in multilingual and code-mixed contexts.

Keywords

CLIP, Vision Transformer, XLM-Roberta, OpenAI, ViT

1. Introduction

Social media networks such as Twitter(X) and Instagram have been among the most popular communication tools, sources of information, and entertainment over the past few years. Such sites, along with their positive effect, have become sources of spreading malicious content, in particular hate speech, abusive comments, and offensive memes. Hate speech is speech that attacks a person or group based on attributes such as race, religion, ethnic origin, national origin, sex, disability, sexual orientation, or gender identity [1]. Memes, in particular, have become one of the most prevalent forms of communication on social media. The combination of images and text commonly makes the insidious spread of hate speech harder to detect than text alone.

In India, there has been a significant increase in hate memes circulating on Twitter and Instagram, especially around sensitive topics such as politics, religion, racism, caste, and the movie industry. Hate speech has evolved on different topics in the last thirty years. [2]. Political memes have been used as a tool of attack in elections to discredit the opponent, and offensive memes against actors and films have been used in online harassment campaigns. This kind of content often employs sarcasm, satire, and coded language, and thus it is hard to detect by moderation systems.

This is even more difficult in India's multilingual and code-mixed social media world. The memes are usually produced in Hindi, Telugu, Bangla, and Tamil (and other regional languages), making it difficult to recognize offensive material.

Shared tasks have been critical in filling these gaps to enhance the research in hate speech detection. CHiPSAL is a shared task that focuses on detecting hate speech in South Asian languages[3]. The HASOC series of shared tasks, organized annually since 2019, has contributed benchmark datasets and standardized evaluation protocols for multiple languages, including Hindi, German, and English [4]. The later editions of HASOC have been extended to include more languages and social media platforms, and have encouraged research in multilingual, code-mixed, and multimodal hate speech detection.

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

nandamsaisaketh.242it021@nitk.edu.in (S. S. Nandam); m_anandkumar@nitk.edu.in (A. K. Madasamy)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-02-24

CEUR-WS.org/Vol-4173/T9-14.pdf

These shared tasks encourage international participation, offering a collaborative environment where effective models can be developed to cope with complexity in the real world. [5, 6, 7, 8]. Moreover, other tracks like TRAC (Trolling, Aggression and Cyberbullying), FIRE shared tasks on offensive language detection, and SemEval tasks [9] on abusive language have all added to the capability of the research community to combat the harmful online material. [10]Introduces a multimodal dataset of Bengali memes combining text and images for abusive content classification, facilitating research on hate and offensive meme detection in low-resource settings.[11] Proposes a modular system combining detection, translation, and mitigation to identify and reduce hate intensity in social media texts across multiple Indic languages.[12] Analyzes results from shared tasks to highlight challenges and performance trends in hate speech detection for low-resource languages, emphasizing data scarcity and model generalization issues.

Hasoc meme 2025 [13]shared task introduced tasks for hate speech and offensive content identification in memes in four Indian languages (Bangla, Hindi, Gujarati, and Bodo). The task is divided into four parts: analyzing the multimodal meme data to detect abuse, assess vulgarity and sarcasm, and assign sentiment labels.

- Sentiment Detection: Classify memes as positive, neutral, or negative.
- Sarcasm Detection: Identify whether a meme is sarcastic or non-sarcastic.
- Vulgarity Detection: Determine whether a meme is vulgar or not vulgar.
- Abuse Detection: Detect whether a meme is abusive or non-abusive.

2. Methodology

This task involves developing a model that takes an input image and its corresponding OCR text, leveraging both visual and textual features to generate an output label. The methodology framework draws from recent advances in vision-language models, particularly the success of contrastive learning approaches that align visual and textual representations in a multimodal task.

2.1. CLIP-based Architecture

This architecture adopts a late fusion strategy where the system processes visual and textual features independently and then combines them for final classification. For visual feature extraction, we use OpenAI's CLIP (openai/clip-vit-base-patch32) which is trained with contrastive learning on a large corpus of publicly available image-caption datasets. For textual encoding, we apply a multilingual Sentence transformer (sentence-transformers/clip-ViT-B-32-multilingual-v1) to effectively handle code-mixed OCR text. The choice of this particular text encoder ensures compatibility with the visual encoder's embedding space while maintaining the ability to process the linguistic complexity inherent in multilingual social media content.

Visual processing starts with CLIP's established pipeline, involving image resizing to 224x224 pixels, normalizing the pixel value in the range [-1,1], and converting the result into a tensor. The visual processing pipeline uses OpenAI's CLIP Vision Transformer (ViT-Base-Patch32) to segment input images into 32x32 patches and process them through a multi-headed self-attention mechanism. Vit model generates a 768-dimensional embedding. To align with the textual embeddings, we project the visual embedding from 768 to 512 dimensions through a learnable projection layer. We freeze the masked self-attention transformer of CLIP (the text encoder) during training and fine-tune only the vision model and the projection layer. We preprocess the OCR-extracted text by removing punctuation and converting English tokens to lowercase. After cleaning, we encode the text with a multilingual Sentence Transformer, which produces 512-dimensional contextualized embeddings that capture semantic meaning and cross-lingual relationships. The fusion mechanism concatenates the 512-dimensional visual embeddings with 512-dimensional textual embeddings to create a comprehensive 1024-dimensional multimodal representation. This concatenated representation is then processed through a feedforward neural network classifier designed with regularization techniques, including batch normalization,

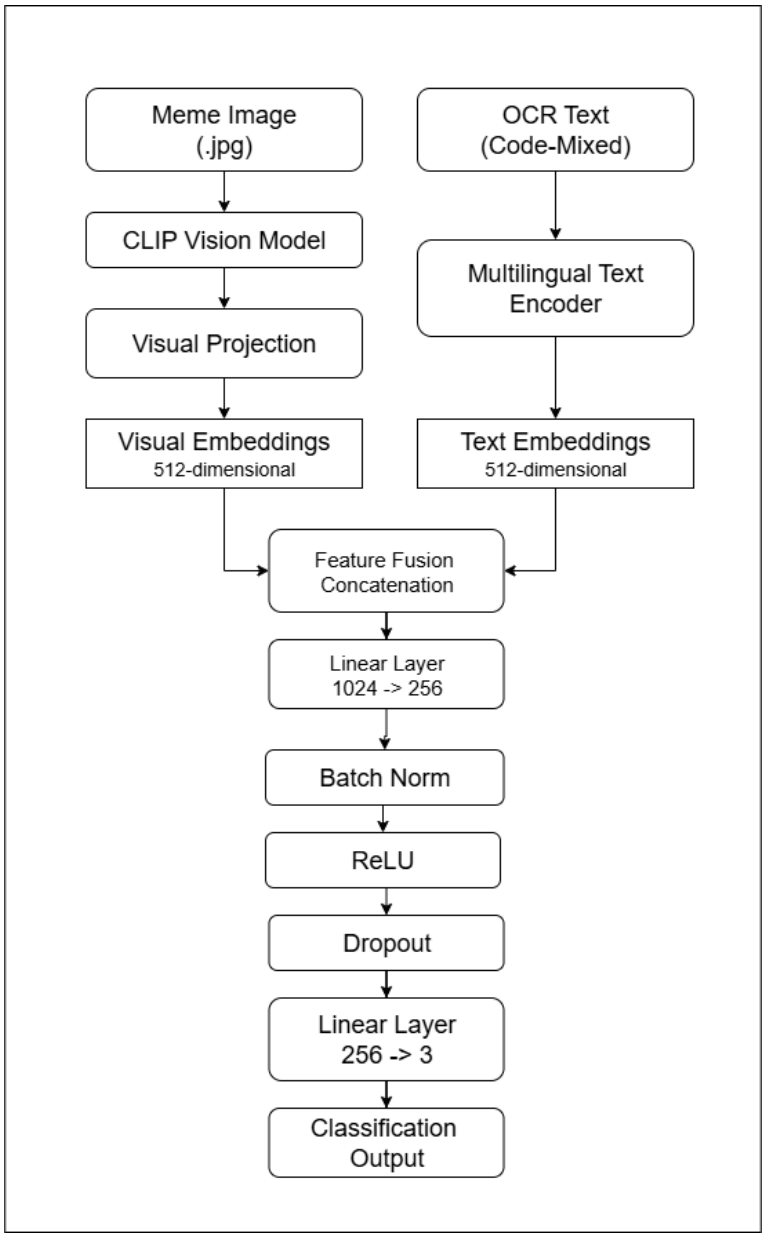

Figure 1: CLIP-based Mulitmodal Architecture combining visual and textual embeddings for meme classification

dropout layer, and residual connections to prevent overfitting. The classifier outputs the final label prediction. The model architecture is shown in Figure 1. We experimented with this architecture using two approaches: (i) fine-tuning both pretrained models jointly, and (ii) freezing the text encoder while fine-tuning only the visual encoder.

2.2. Vit + XLM-Roberta Architecture

In this architecture, we use Visual Transformer (ViT-B-16-plus-240) as the visual encoder and XLM-Roberta (M-CLIP/XLM-Roberta-Large-Vit-B-16Plus) as the textual encoder for feature extraction.

For the visual pipeline, we preprocess each meme image by resizing and normalizing it, then pass it through a vision transformer encoder, which produces a 640-dimensional embedding. For the textual pipeline, we apply similar preprocessing steps, followed by encoding with a multilingual text encoder. This text encoder generates 640-dimensional embeddings that capture both semantic and cross-lingual relationships.

We then concatenate the visual and textual embeddings to form a unified 1280-dimensional multimodal representation. This representation passes through a feedforward classifier consisting of a linear projection layer, batch normalization, ReLU activation, dropout, and a final linear layer to produce task-specific labels. We experimented with this architecture using two approaches: (i) fine-tuning both pretrained models jointly, and (ii) freezing the text encoder while fine-tuning only the visual encoder.

3. Experimental Setup

3.1. Dataset Statistics

The data set provided comprises a multilingual collection of meme images in Four languages: Hindi, Gujarati, Bangla, and Bodo, totaling 5,101 samples, with sizes ranging from 378 (Bodo) to 2,693 (Bangla). Each language data the set is divided into training sets (70%), validation sets (15%), and developement set (15%), ensuring a balanced representation for model training and evaluation. The dataset statistics are mention in Table 1. Details on the dataset can be found in [13], the HASOC 2025 overview paper.

Table 1
Dataset Statistics of HASOC Meme 2025

Language	Dataset Size	Train	Validation	Development	Test
Hindi	1141	798	171	172	770
Gujarati	889	623	133	133	604
Bangla	2693	1885	404	404	1821
Bodo	378	264	57	57	254

3.2. Implementation Details

All experiments were conducted on the Kaggle platform using an NVIDIA Tesla T4 GPU for accelerated model training. AdamW Optimizer is employed for parameter updates, with gradient clipping applied at a maximum norm of 1.0 to stabilize training and prevent exploding gradients. All hyperparameter settings are mentioned in Table 2.

Table 2
Hyperparameter Settings

Parameter	Value
Batch Size	8
Epochs	3
Classifier Dropout Rate	0.6
Optimizer	AdamW
Weight Decay	0.1
Loss Function	Cross Entropy Loss

For learning rate configurations, careful tuning was performed for each model component mentioned in Table 3. We use the same set of hyperparameters consistently across all three models.

- Vision Model (2e-6): An extremely low learning rate was employed to fine-tune the pre-trained CLIP vision backbone gently, leaving intact its overall visual knowledge, but adapting it to the idiosyncrasies of memes.
- Visual Projection Layer (3e-5): A marginally higher rate was selected to enable the projection layer to match the image and text embeddings effectively so that multimodal fusion can be achieved without destabilizing the vision model.

Table 3
Learning Rate of Model Components

Model Components	Learning Rate
Vision Encoder	2e-6
Visual Projection	3e-5
Text Encoder	5e-7
Classifier	5e-5

- Text Encoder (5e-7): The low learning rate was used to avoid overfitting the pre-trained multilingual SentenceTransformer and retain a good representation of the multilingual text.
- Classifier (5e-5): A moderate learning rate allowed the classifier to successfully learn task-specific decision boundaries (e.g., sarcasm detection) in the fused embeddings, minimizing the chances of overfitting to the small dataset.

4. Experimental Results

We use the Micro F1 score to test our models, which is in line with the official evaluation protocol for the shared task. For final leaderboard placement, team rankings are determined by the average Macro F1 score across four subtasks: sentiment classification, sarcasm detection, vulgarity detection, and abuse detection. Table 4 presents the Micro F1 scores of different model configurations across four languages: Hindi, Bangla, Gujarati, and Bodo.

Table 4
Performance results of multimodal architectures across four languages. The visual encoder is fine-tuned in all cases, while the text encoder is either fine-tuned or frozen during training.

Language	Visual Encoder	Text Encoder	Micro F1
Bangla	CLIP-ViT-B/32	CLIP-Multi	**0.61034**
	CLIP-ViT-B/32	CLIP-Multi (Frozen)	0.59588
	ViT-B/16+240	XLM-Roberta	0.45215
	ViT-B/16+240	XLM-Roberta (Frozen)	0.54503
Hindi	CLIP-ViT-B/32	CLIP-Multi	0.56627
	CLIP-ViT-B/32	CLIP-Multi (Frozen)	**0.58468**
	ViT-B/16+240	XLM-Roberta	0.49585
	ViT-B/16+240	XLM-Roberta (Frozen)	0.57225
Gujarati	CLIP-ViT-B/32	CLIP-Multi	**0.61715**
	CLIP-ViT-B/32	CLIP-Multi (Frozen)	0.61239
	ViT-B/16+240	XLM-Roberta	0.55194
	ViT-B/16+240	XLM-Roberta (Frozen)	0.60565
Bodo	ViT-B/16+240	XLM-Roberta	0.59336
	ViT-B/16+240	XLM-Roberta (Frozen)	**0.60393**

- Bangla: The highest score was obtained with the unfrozen CLIP-ViT/CLIP-multi model (0.61034). Despite the relatively small gap between frozen and unfrozen settings, our approach demonstrated strong adaptability to Bangla, resulting in a 4th out of 17 teams.
- Hindi: The frozen CLIP-ViT/CLIP-multi model achieved the best performance with a Micro F1 of 0.58468, outperforming both its unfrozen counterpart and the ViT/XLM-Roberta baseline. Our system ranked 5th out of 18 teams for this language.

- Gujarati: Similar to Bangla, the unfrozen CLIP-ViT/CLIP-multi achieved the best Micro F1 (0.61715). The CLIP framework consistently outperformed the ViT/XLM-Roberta baseline and secured a 4th out of 15 teams.
- Bodo: Since the CLIP-multilingual model is pre-trained on 56 languages and Bodo is not included among them, only the ViT/XLM-Roberta model was fine-tuned for this language. Frozen text encoder method achieved a Micro F1 of 0.60393, positioned our team 4th out of 15 teams.

In short, the CLIP-ViT/CLIP-multi model showed strong multilingual generalization abilities, particularly for Bangla and Gujarati, whose best results were achieved by full fine-tuning. Conversely, for Hindi, the frozen text encoder of CLIP framework gave better results, indicating that fine-tuning may not always be beneficial in low-resource multilingual contexts. In Bodo, where CLIP training coverage was absent, the ViT-XLMR baseline proved a competitive choice. Our consistent top-5 rankings across all four languages highlight the effectiveness of multimodal CLIP-based models for multilingual meme classification and related subtasks.

5. Conclusion

Based on the experimental findings, combining visual and textual features in a common multimodal framework proved to be much better regarding classification results in all the tasks. The models based on the CLIP ViT vision backbone and the multilingual text encoders demonstrate high performance when it comes to capturing the fine details of the interaction between imagery and language, which is especially relevant when it comes to picking up the subtle categories like sarcasm, abuse, and vulgarity.

Differential learning rates were found useful in refining the parameters of significant pre-trained components without affecting the generalization capacity of such components. Slower learning rates on the vision and text encoders preserved the semantic richness they had been pre-trained on. However, faster rates on the projection and classification layers allowed more efficient adaptation to the features in the dataset.

Overall, the proposed multimodal architecture is a powerful and flexible framework for multilingual meme classification. Our system got top-5 rankings in all four languages: 5 out of 18 in Hindi, 4 out of 17 in Bangla, 4 out of 15 in Gujarati, and 4 out of 15 in Bodo. Future work may further consider more sophisticated vision backbones and adaptive fusion mechanisms to improve performance in various cultures and languages.

Declaration on Generative AI

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

References

[1] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, O. Frieder, Hate speech detection: Challenges and solutions, PloS one 14 (2019) e0221152.

[2] A. Tontodimamma, E. Nissi, A. Sarra, L. Fontanella, Thirty years of research into hate speech: topics of interest and their evolution, Scientometrics 126 (2021) 157–179.

[3] K. Sarveswaran, A. Vaidya, B. K. Bal, S. Shams, S. Thapa, Proceedings of the first workshop on challenges in processing south asian languages (chipsal 2025), in: Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025), 2025.

[4] 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, 2019, pp. 14–17.

[5] T. Mandl, S. Modha, A. Kumar M, B. R. Chakravarthi, Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german, in: Proceedings of the 12th annual meeting of the forum for information retrieval evaluation, 2020, pp. 29–32.

[6] S. Modha, T. Mandl, G. K. Shahi, H. Madhu, S. Satapara, T. Ranasinghe, M. Zampieri, Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages and conversational hate speech, in: Proceedings of the 13th annual meeting of the forum for information retrieval evaluation, 2021, pp. 1–3.

[7] T. Ranasinghe, K. North, D. Premasiri, M. Zampieri, Overview of the hasoc subtrack at fire 2022: offensive language identification in marathi, arXiv preprint arXiv:2211.10163 (2022).

[8] S. Satapara, S. Masud, H. Madhu, M. A. Khan, M. S. Akhtar, T. Chakraborty, S. Modha, T. Mandl, Overview of the hasoc subtracks at fire 2023: Detection of hate spans and conversational hate-speech, in: Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation, 2023, pp. 10–12.

[9] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti, Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter, in: Proceedings of the 13th international workshop on semantic evaluation, 2019, pp. 54–63.

[10] 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.

[11] 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.

[12] 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.

[13] 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.