Hateful and Offensive Meme Detection in Multimodal Memes Dataset for Indo-Aryan Languages

Aarsh Sarvaiya[1], Tripti Kumari[2,*] and Ayan Das[2]

[1]Sardar Vallabhbhai National Institute of Technology, Surat, Gujarat, 395007, India
[2]Department of Computer Science and Engineering,
Indian Institute of Technology Dhanbad, Jharkhand, 826004, India

Abstract

This paper presents the system submitted by the team from IIT Dhanbad in the FIRE IRSE HASOC-2025 shared task on automatically identifying hateful and offensive memes in four Indian languages—Gujarati, Hindi, Bengali, and Bodo. Here, we developed a multimodal deep learning framework for five subtasks of Sentiment, Sarcasm, Vulgarity, Abuse, and Target Communities to automatically predict the hateful and offensive memes. we have used a task-specific strategy in which a different multimodal model handles each label. In addition to methods like gated attention and weighted losses, we employed a variety of text encoders (XLM-R, mBERT, MuRIL, BanglaBERT) and image encoders (EfficientNet, DenseNet, VGG19, ResNet). Extensive experiments show that task-specific pipelines consistently provide better performance when combined using ensemble methods. These results demonstrate the value of modular multimodal models in addressing the complex, varied, and frequently implicit character of hate in memes. In the official assessment, the final leader board's best macro F1-scores on the Kaggle platform for Hindi, Gujarati, Bangla, and Bodo were 0.5741 (6^{th} rank), 0.5887 (8^{th} rank), 0.5720 (8^{th} rank), and 0.5818 (7^{th} rank), respectively.

Keywords

Hateful and offensive memes detection, Hindi, Gujarati, Bengali, Bodo, Multi-modal memes, Embeddings, Neural networks, NLP, Transformers, Indo-Aryan languages, Text/image-based transformers, Meme classification.

1. Introduction

A meme is a joke, cultural concept, or trend that is conveyed through text, images, or videos and quickly gains traction on social media. Over the past ten years, social media's influence has fundamentally changed how people express themselves, share their thoughts, and participate in public discussions. Memes have emerged as one of the most widely used and influential online communication tools. They combine images with just enough text to tell a story or make a joke; they are brief, visually appealing, and frequently humorous. But more often than not, the same format is also being twisted into a platform for hate speech, sarcasm, targeted abuse, and vulgarity[1, 2]. In India, where memes are created in a variety of languages, dialects, transliterations, slang, and emojis, moderation is particularly difficult [3]. This presents a serious challenge to maintaining the security of digital platforms.

If platforms are to safeguard users and stop the spread of harmful content, they must be able to flag hateful memes with speed and accuracy. However, the task is much more difficult than it seems. Because they combine text and images to convey meaning, memes are by nature multimodal. Optical Character Recognition (OCR) is frequently used to extract text from images; however, in practice, this results in outputs that are noisy, incomplete, or distorted because of stylized fonts or poor image quality [4]. The problem becomes evident when one considers the unbalanced nature of datasets and the fact that hate speech on the internet frequently goes unnoticed due to irony, humor, or coded language [5, 6]. Building multimodal models that can truly handle the messiness and variety of Indic languages is still a challenge, despite the fact that deep learning has given us powerful tools for understanding text and images separately [7].

Forum for Information Retrieval Evaluation, December 17-20, 2025, India
*Corresponding author.
sarvaiya.aarsh438@gmail.com (A. Sarvaiya); 22dr0264@iitism.ac.in (T. Kumari); ayandas@iitism.ac.in (A. Das)

Previous studies have demonstrated the effectiveness of Transformer-based models for handling text [8, 9] and Convolutional Neural Networks (CNNs) for extracting image features [10]. However, it rarely works right out of the box to plug these into meme datasets. The signals are jumbled, occasionally contradictory, and contain either incomplete or deceptive text and images. Models that can jointly reason over both modalities without being distracted by noise are required [11].

In this study, we investigate three methods for classifying memes in five different subtasks: target communities, sarcasm, vulgarity, abuse, and sentiment. In the first, shared encoders try to learn everything together in a unified multimodal pipeline. The second, a hierarchical transformer fusion model (HIT-FAME), presents attention-based mechanisms that facilitate better alignment of text and image features [12]. The third is based on a task-specific modular pipeline wherein individual multimodal models are trained for every label and subsequently integrated through ensemble techniques [13]. We test a number of image encoders (EfficientNet [14], DenseNet [15], VGG19 [16]) and text encoders (mBERT [8], XLM-R [9], MuRIL [17]) across all strategies. We also use preprocessing techniques like emoji mapping, profanity masking, and transliteration handling to overcome the peculiarities of Indic memes [18].

This paper makes two contributions. First, we demonstrate that ensemble strategies in conjunction with task-specific modular pipelines consistently outperform unified models, especially when dealing with multilingual meme languages and noisy OCR text. In order to shed light on the ways preprocessing decisions, encoder selection, and fusion strategies affect performance, we secondly present a benchmark framework for multimodal hate speech detection in Indic settings [19].

2. Literature Survey

Three interconnected insights from recent studies on multimodal hateful, abusive, and subjective content—particularly memes that combine images with text overlayed or associated with them—directly inform the HASOC 2025 challenge[20]. The first is that evaluation design and datasets are important. The Hateful Memes benchmark showed that models can exploit unimodal artifacts unless datasets explicitly include examples of adversarial or "benign confounders," which are examples that require real image–text compositional reasoning [2]. Likewise, Memotion-3's Hinglish/code-mixed dataset [21, 22] supports the Memotion series' useful annotation schemas for sentiment, humour/sarcasm, and offensiveness in memes. It also demonstrated the importance (and challenge) of gathering culturally diverse memes. Second, large-scale vision–language pretraining and modular multimodal approaches have created strong, robust visual representations that transfer well to downstream tasks[23], while more recent work (e.g., BLIP-2, Flamingo, and visual-instruction tuning) has made it possible to create flexible, instruction-capable multimodal models that work well in zero-shot and few-shot scenarios [24, 25, 26]. Although these developments shift the failure modes (e.g., over-reliance on pretraining biases), they also make it possible to construct HASOC systems that combine powerful language reasoning with strong visual embeddings. Third, task-specific work reveals unique challenges for target identification, abuse, vulgarity, sarcasm, and sarcasm: (a) ironic intent and sarcasm frequently stem from subtle visual-textual incongruity, necessitating explicit modeling of mismatch signals and cross-modal fusion; (b) vulgarity and profanity are not interchangeable with targeted hate—practical moderation systems must differentiate between profanity and abusive content that targets protected groups; and (c) target identification frequently calls for cultural and background knowledge beyond surface tokens, which is particularly acute in non-English and code-mixed memes [21, 5, 22]. There are still several methodological and assessment gaps in these strands. Even though it is evident that code-mixing and regional references have a significant impact on both labels and model performance, dataset creation still faces challenges with annotation subjectivity (low agreement on sarcasm, humor, and intensity) and many benchmarks are still English-centric [27, 2, 22]. Although cross-attention transformers and contrastive vision-language encoders offer robust architectural foundations, they fall short in addressing multimodal compositionality and robustness to adversarial confounders. It has been demonstrated that combining OCR-derived text signals with region-aware visual features and

contrastive or contrastive-finetuning objectives is beneficial but insufficient [2, 24]. Lastly, academic benchmarks under-develop evaluation metrics and deployment-oriented criteria (such as false-positive risk for protected groups, interpretability of decisions, and cultural calibration), which are essential for moderation in the real world.

Overall, previous work offers both useful toolkits (OCR pipelines, multimodal transformers, CLIP-style visual embeddings, and instruction-tuned multimodal LLMs) and warnings (cultural bias, unimodal shortcuts, and annotation subjectivity). The literature recommends that for HASOC 2025 [20, 28], evaluation protocols that measure cross-modal compositionality and cultural robustness instead of just aggregate accuracy, multilingual/code-mixed coverage, explicit separation of profanity vs. targeted abuse, and adversarial-aware dataset design. Following these paths will bring model evaluation and practical requirements for multimodal meme moderation closer together.

3. Experiment design

This section provides a thorough explanation of the suggested systems created for predicting hate speech detection experiments that have been conducted across the various Indo-Aryan languages, including Bengali, Hindi, Gujarati, and Bodo, to solve the four distinct tasks. The proposed framework has been shown in the figure 3. We have given more detail about our methodology, including word embeddings and neural transformer architecture training, in the upcoming section.

3.1. Description of dataset and Tasks

FIRE HASOC-2025 organizers provided the datasets[1], which were primarily two training and testing datasets with images (JPEG) of memes gathered as depicted in Figure 1, arranged by language, and a.csv file with the image id. As illustrated in Figure 2, represents the overall true labels with noisy and uncleaned OCR data extracted from the images they provided, including languages such as Bengali, Hindi, Bodo, and Gujarati.

Figure 1: Sample of memes dataset of Bengali language

Id	Sentiment	Sarcasm	Vulgarity	Abuse	Target	OCR
Gujarati_image_1618.jpg	Positive	Sarcastic	Vulgar	Abusive	Group	àªœà«‡àª® àª¬àª¨àª¾àªµà«•àª, àª¤àª¾àª«à«•...
Hindi_image_0432.jpg	Negative	Non-Sarcastic	Vulgar	Abusive	Individual	à¤¦à¥‡àª¶ à¤•à¤¾ à¤•à¤¾à¤¨à¥‚à¤¨...
Bengali_image_0110.jpg	Neutral	Non-Sarcastic	Non-Vulgar	Non-Abusive	Others	à¦•à¦‡ à¦• à¦¹à§Ÿ à¦¨à¦¾ ...
Bodo_image_0091.jpg	Negative	Sarcastic	Non-Vulgar	Abusive	Gender	à¦†à¦‰ à¦²à¦¾à¦–à¦¾à¦¯à¦¼à¦....

Figure 2: Description of the HASOC shared task-2025 memes dataset across Indo-Aryan languages

[1]https://hasocfire.github.io/hasoc/2025/dataset.html

3.2. Data preprocessing

There were emojis, profanities, informal structures, and noise in the meme OCR text. The following procedures were used to apply a strong cleaning function consistently across all languages:

- **Emoji Replacement:** To translate commonly used emojis into textual representations, a carefully chosen emoji-to-word mapping was used.
- **Lower casing and Numeric Normalization:** To maintain semantic intent, all English characters were lowercased, and a <num> placeholder was used in place of numerical values.
- **Removal of Unwanted Tokens:** Regex patterns were used to eliminate URLs, hashtags, mentions, and mixed-script artifacts.
- **Mixed-Script Filtering:**To get rid of OCR-related errors, words with both Latin and Indic script characters were removed.
- **Stop Word Removal:** Both native and Romanized tokens were included in the unique list of stop words used by each language.
- **Profanity Masking:** For every language (Roman + native script), a longer list of offensive terms was kept up to date, and matches were concealed by using the <PROFANITY> token[2]. The list of offensive words for each language, which is prepared for profane words:
  Bengali= 'chod', 'chodon', 'chudi', 'gandu', 'bal', 'randi', 'khanki', 'madarchod', 'banchod', 'bokachoda', 'shala', 'shali', 'haramzada'
  Hindi = 'bhenchod', 'madarchod', 'gaandu', 'gandu', 'chutiya', 'chutiye', 'bhosdike', 'bhosdiwale', 'bsdk', 'mc', 'bc', 'randi', 'kutta', 'kamine', 'kamina', 'harami', 'haramkhor', 'lund', 'loda', 'lauda', 'jhaant', 'gote', 'tatte', 'choot', 'behenchod', 'maaki', 'gaand', 'bhadwa', 'lavde'
  Gujarati = 'Chodhru', 'chodkanya', 'Chodu', 'Chodu Bhagat', 'Choidi Rand', 'Choot Marina', 'Gaand Ma Ghal', 'gaandina', 'gandina', 'gand ma ghali de', 'Gand Maraav', 'Gand nu kaanu', 'gandi chodina', 'Gandi Gaan', 'gando', 'Gando salo', 'Ghel Sappo', 'Ghelchoydi', 'Halki Raand na', 'hopa', 'Jhaanth', 'Jhadhino!', 'Lukkha Loda', 'luli', 'Lund Bhagat', 'mota jijah', 'Moti Gaand', 'nago choido', 'Nakhodiya', 'taari gaand ma dando ghalu'
  Bodo = 'abu gidir', 'Bazari', 'Bima khoigra', 'Bima khoygra', 'Bimani fishai', 'Bwitali', 'bwithala', 'Cfa swlangra', 'Hinjao maogra (khoigra)', 'khilama', 'Khoynai', 'Lwdwi', 'Nwma shifa', 'Nwmani abu', 'Sifa gudung', 'Sifa jagra', 'sifa jagra', 'Sifa swlagra', 'Sikwmwn', 'swima ni pisa', 'tapa'
- **Whitespace and Punctuation Normalization:** Excessive punctuation was eliminated, and one space was used in place of several. This procedure made sure that every sample had clear, comprehensible text devoid of harmful or unnecessary tokens.

3.3. Text encoder

Particularly for Indic memes, where the textual component may be embedded in noisy, stylized, or code-mixed OCR outputs, textual comprehension is an essential part of multimodal hate speech detection. We experimented with and used both pretrained Transformer-based encoders and a custom static embedding approach (HIT-BPE) to efficiently process such content across multiple languages. Task-wise performance was used to select or combine each encoder. Before being fused with the visual modality, all encoders convert the final cleaned and pre-processed text into a fixed-length vector representation. mbert, XLM, and MuRIL have all been employed as text encoders.

3.4. Image encoder

We employed deep convolutional neural networks (CNNs) for image encoding in order to supplement the textual stream. These use visual cues such as facial expressions, symbols, gestures, or embedded text to extract high-level semantic and spatial features from meme images that may suggest hate, sarcasm,

[2]Warning: This paper contains certain offensive or potentially upsetting content, which is included only due to the nature of the work and could not be avoided

Model	Output Dim	Used For Tasks	Strengths
xlm-roberta-base	768	vulgar, sarcasm	Robust multilingual, cross-lingual
mBERT	768	sentiment, target comm.	Transferable & widely used, multilingual
Google MuRIL	768	vulgar, abuse	Indian language focused, efficient NLU
BanglaBERT	768	abuse, sentiment	Bengali optimized, high accuracy

Table 1
Comparison of Models for Multimodal Meme across different Classification Tasks

sentiment, or abusive context. A resized and normalized RGB image is converted by each image encoder into a fixed dimensional feature vector, usually 512, which is then projected and fused with the textual representation in the multimodal fusion module. These models have been used, including DenseNet-161, ResNet-152, EfficientNet-B7, and VGG19-BN (Batch Normalized VGG19).

Model	Output Dim	Used For Tasks	Strengths
VGG19-BN	25088	Sarcasm, Abuse, Vulgar	Simpler, deeper CNN
DenseNet-161	2208	Sentiment, Target	Dense connections, feature reuse
ResNet-152	2048	Vulgar, Sarcasm	Deep + stable gradients
EfficientNet-B7	2560	Target, Sentiment	Efficient and high-performing

Table 2
Comparison of CNN Models for Meme Classification Tasks

3.5. Fusion Mechanism and Classification Head

All five classification tasks can be powered by our fusion strategy, which combines the two modality streams—text and image—into a single joint representation. There are three stages to it:

1. Modality Concatenation

- A single 1024-dimensional vector is created by concatenating the 512-dimensional text and 512-dimensional image embeddings.
- By preserving all of the data from both encoders, this raw concatenation enables the model to focus on the features that are most predictive for each sample.

2. Gated Fusion

- A sigmoid activation is performed after a single linear transformation to compute a learnable ”gate” vector of the same size (1024).
- The concatenated features are multiplied element-wise by this gate, which functions similarly to a soft mask.
- In practice, the gate learns to up- or down-weight particular joint features, such as highlighting image cues when the image itself is offensive or text cues when the meme's humor is text-centric.

3. Joint MLP and Task Heads

- A three-layer MLP (1024⟶512⟶256⟶128) with ReLU activations and 30% dropout at each layer is used to process the gated 1024-dim vector.
- Higher-order interactions between the two modalities are learned by this MLP.
- The 128-dim representation is then projected to the task's label space by a specialized linear head for each of the five tasks (2 classes for Abuse/Vulgarity/Sarcasm, 3 classes for Sentiment, and 8 classes for Target communities).
- The multi-class heads use softmax, while the binary heads use sigmoid activations at inference.

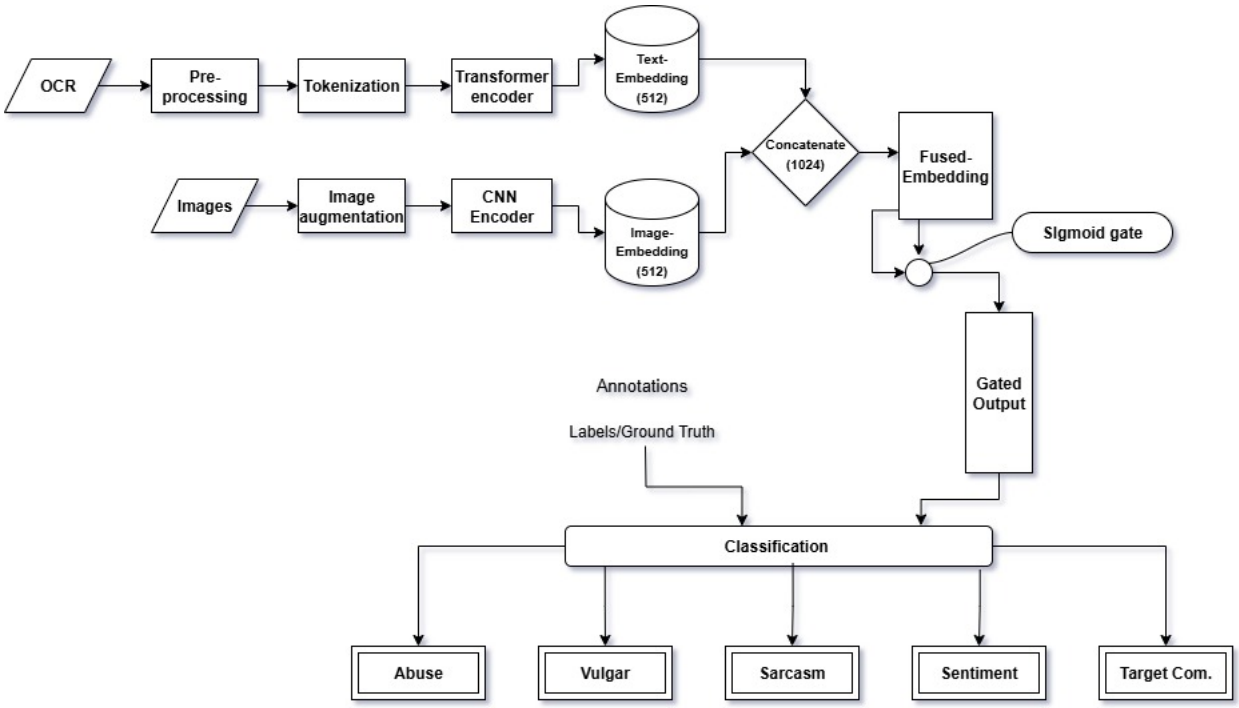

Figure 3: Flow diagram of proposed framework to automatically predict the hateful and offensive memes across the shared tasks of different Indo-Aryan languages

4. Experimental results and Analysis

We present the results of the FIRE HASOC-2025 task, which entails evaluating multimodal data (text and images) in order to identify targeted communities, detect abuse, evaluate sarcasm and vulgarity, and assign sentiment labels.

4.1. Dataset

The FIRE HASOC-2025 organizers initially provided training datasets containing meme images (JPEG) and corresponding CSV files with image IDs. These four individual datasets covered the languages—Bengali, Hindi, Bodo, and Gujarati—and included noisy, uncleaned OCR text extracted from the images. We split the training data into 80:20 ratios for training and testing to evaluate the best-performing model for each language across the different tasks described in the Table 3

4.2. Tasks

4.2.1. Abuse Detection

In Bengali (BanglaBERT+VGG19, F1 = 0.7938) and Gujarati (Muril+VGG19, F1 = 0.7973) perform the best, as indicated in Tables 6 and 4. According to tables 5 and 7, Hindi and Bodo perform somewhat worse (about 0.74–0.76 F1). The effectiveness of the pretrained models (BanglaBERT for Bengali and Muril for Gujarati/Hindi) in conjunction with CNN backbones emphasizes the significance of language-adapted models for abuse detection.

4.2.2. Vulgarity Detection

In Bengali has the highest score (XLM-R + DenseNet161, F1 = 0.8496), as indicated in tables 6 and 4. Gujarati comes in second (0.8416), as indicated in tables 4. As indicated in tables 5 and 7, respectively, Hindi and Bodo continue to fall within the mid-0.73–0.76 range. Here, as illustrated in figure 4, we discovered that the robust cross-lingual models, such as XLM-R, when combined with deeper CNNs

Table 3
The initially split training and test sets from the training data were used to select the best performing models across different languages and shared tasks

Language	Train Samples	Test Samples	Task	Model Combination
Gujarati	711	178	Abuse	MuRIL + VGG19
			Vulgar	XLM-R + VGG19
			Sarcasm	XLM-R + DenseNet161
			Sentiment	MuRIL + VGG19
			Target Comm.	MuRIL + VGG19
Hindi	912	229	Abuse	XLM-R + EfficientNetB7
			Vulgar	XLM-R + VGG19
			Sarcasm	MuRIL + VGG19
			Sentiment	XLM-R + VGG19
			Target Comm.	XLM-R + VGG19
Bangla	2154	539	Abuse	BanglaBERT + VGG19
			Vulgar	XLM-R + DenseNet161
			Sarcasm	BanglaBERT + VGG19
			Sentiment	BanglaBERT + VGG19
			Target Comm.	mBERT-uncased + EfficientNetB7
Bodo	302	76	Abuse	mBERT + EfficientNetB7
			Vulgar	MuRIL + VGG19
			Sarcasm	XLM-R + VGG19
			Sentiment	XLM-R + VGG19
			Target Comm.	mBERT + VGG19

(DenseNet, VGG19), perform exceptionally well at vulgar detection, indicating generalizability across languages.

4.2.3. Sarcasm Detection

As indicated in Tables 6 and 7, respectively, Bengali once again demonstrates promising results with BanglaBERT+VGG19 (F1 = 0.8229), while Gujarati (0.6793) and Hindi (0.6801) trail behind. When available, language-specific models perform better than multilingual ones, but Sarcasm is still a difficult task in low-resource environments.

4.2.4. Sentiment Classification

The F1-scores for all languages are generally lower, ranging from 0.41 for Bodo to 0.67 for Bengali, as indicated in 7 and 6, respectively. Bengali benefits from BanglaBERT (0.6720), while Gujarati and Hindi both struggle (0.49–0.64). Here, we discovered that multimodality and subtle cues make sentiment detection in memes extremely difficult; results are marginally improved by language-specific embeddings.

4.2.5. Targeted Community Detection

In this task, the extremely low F1-scores (0.23–0.42), this is the most difficult task in all languages. Bengali (mBERT+EfficientNetB7, F1 = 0.4215) was the best result, as indicated in table 6, and Hindi (0.2319) performs the worst among the other languages. Here, we found that the lack of data, implicit hate speech, and the current models' poor contextual awareness make it difficult to identify target communities.

The performance of various models on a range of classification tasks is shown in the figure 4. Here, the BanglaBERT + VGG19 and XLM-R + VGG19 models obtain the highest F1 scores, especially for the Abuse and Sarcasm tasks. On the other hand, several models display F1 scores below 0.5, suggesting that tasks such as Target Comm. are extremely difficult.

A direct comparison of the F1 scores for each task across the languages is shown in Figure 5. With the highest F1 score for the Sarcasm and Sentiment tasks, this figure demonstrates Bengali's consistently strong performance. The Sentiment and Target Comm. tasks have the lowest F1 scores for the Bodo language, despite its strong performance on other tasks.

Figure 5 and 6 illustrates the differences in performance for each language on various tasks. For the Abuse and Vulgar tasks, all four languages—Gujarati, Hindi, Bengali, and Bodo—display high F1 scores. However, F1 scores for the Sentiment and Target Comm. tasks are significantly lower than those for the other languages, indicating a significant decline in performance for all languages.

Task	Model Combination	F1-score	Precision	Recall
Abuse	Muril+VGG19	0.7973	0.7624	0.8231
Vulgar	XLM-R+VGG19	0.8416	0.8477	0.8272
Sarcasm	XLM-R+Densenet161	0.6793	0.6551	0.7119
Sentiment	Muril+VGG19	0.6487	0.6340	0.6520
Target Comm.	Muril+VGG19	0.3233	0.3019	0.2869

Table 4
Performance metrics (F1-score, Precision, Recall) of shared tasks in the Gujarati language.

Task	Model Combination	F1-score	Precision	Recall
Abuse	XLM-R+EfficientNetB7	0.7646	0.7550	0.7535
Vulgar	XLM-R+VGG19	0.7597	0.7675	0.7386
Sarcasm	Muril+VGG19	0.6801	0.6676	0.6702
Sentiment	XLM-R+VGG19	0.4925	0.5136	0.4913
Target Comm.	XLM-R+VGG19	0.2319	0.4255	0.2206

Table 5
Performance metrics (F1-score, Precision, Recall) of shared tasks in the Hindi language.

Task	Model Combination	F1-score	Precision	Recall
Abuse	BanglaBERT+VGG19	0.7938	0.8305	0.7768
Vulgar	XLM-R+Densenet161	0.8496	0.8529	0.8229
Sarcasm	BanglaBERT+VGG19	0.8229	0.8237	0.8346
Sentiment	BanglaBERT+VGG19	0.6720	0.7152	0.6586
Target Comm.	mBERT-uncased+EfficientNetB7	0.4215	0.4091	0.4656

Table 6
Performance metrics (F1-score, Precision, Recall) of shared tasks in Bengali language.

Task	Model Combination	F1-score	Precision	Recall
Abuse	mBERT+EfficientNetB7	0.7426	0.7412	0.7665
Vulgar	Muril+VGG19	0.7364	0.7922	0.7301
Sarcasm	XLM-R+VGG19	0.7964	0.8174	0.7812
Sentiment	XLM-R+VGG19	0.4143	0.3286	0.6115
Target Comm.	mBERT+VGG19	0.3435	0.3376	0.3498

Table 7
Performance metrics (F1-score, Precision, Recall)of shared tasks in Bodo language .

Here, we have used different transformer models because each subtask required distinct linguistic understanding. For example, XLM-R was used for Vulgarity and Sarcasm due to its strong cross-lingual and contextual capabilities. mBERT was used for Sentiment and Target Community tasks as it handles general multilingual contexts well. This task-specific approach improved accuracy since no single model worked best for all subtasks. In short, the Selection of transformer model and their combination

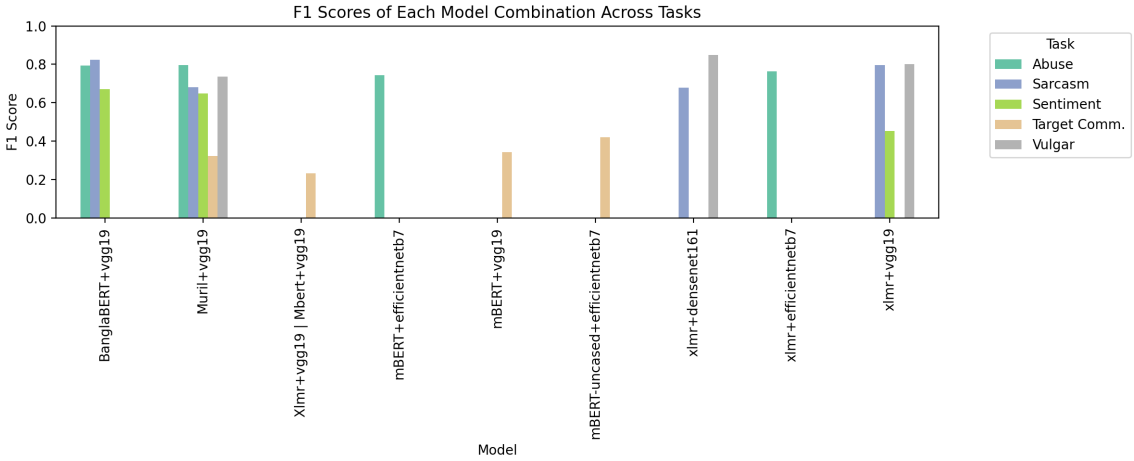

Figure 4: Bar plot of F1- score of each model across the tasks

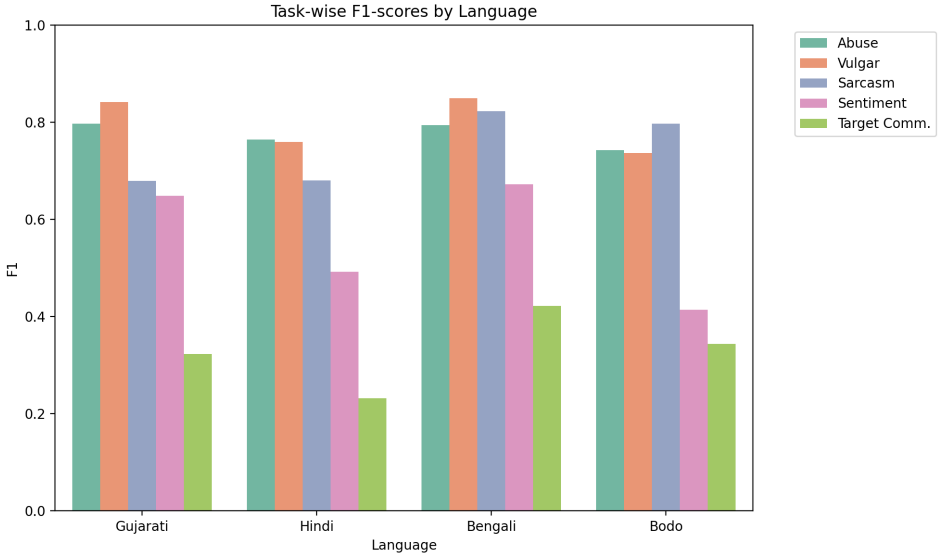

Figure 5: Bar plot of task-wise F1-score by each language

was solely done based on their performance on various specific tasks (Vulgar, Sarcasm, Sentiment, Abuse) and thereby used to boost the overall performance and prediction of the Memes

5. Discussion

When language-specific models (BanglaBERT ([29]), Muril ([17])) are used in conjunction with CNN architectures like VGG19 ([16]) and DenseNet ([15]), the evaluation across Gujarati, Hindi, Bengali, and Bodo shows that abuse and vulgarity detection are comparatively more successful. Sarcasm detection in Bengali and Bodo is still difficult, but it has promise. Given the complexity of multimodal sentiment, sentiment classification produces lower overall scores. With poor performance in all languages, targeted community detection is the most challenging task, highlighting the need for better context modeling and richer datasets. Overall, Bodo's low resource status exposes the shortcomings of the current multilingual approaches, whereas Bengali enjoys the advantages of robust monolingual resources.

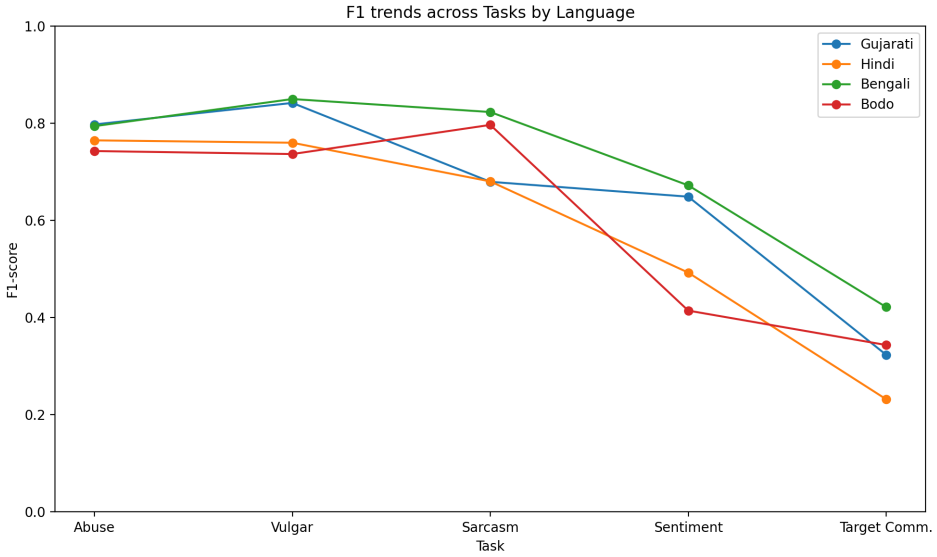

Figure 6: Plots of F1 trends across tasks by language

6. Conclusion

In this paper, we presented a multimodal deep learning framework for five subtasks of Sentiment, Sarcasm, Vulgarity, Abuse, and Target Communities to automatically predict the hateful and offensive memes across the Indo-Aryan languages such as Hindi, Gujarati, Bengali, and Bodo. This work addressed the complex problem of detecting hateful and offensive content in multimodal memes across four Indian languages by leveraging a range of text and image encoders, preprocessing strategies, and multimodal fusion techniques. Our findings show that task-specific modular pipelines, when combined with ensemble strategies, consistently outperform unified multimodal models, particularly in handling noisy OCR text, multilingual variation, and implicit signals of hate. The results highlight that no single model is universally effective across all subtasks, and instead, modularity and flexibility are crucial for robust meme classification. Importantly, our framework not only sets a benchmark for multimodal hate speech detection in Indic contexts but also underscores the need for richer, balanced datasets and context-aware architectures that can capture the subtle interplay of humor, sarcasm, and coded language.

In the future, we will explore integrating external knowledge sources, advanced cross-lingual embeddings, and continual learning approaches to further improve performance in low-resource settings and ensure safer digital environments.

Acknowledgments

The authors acknowledge the Science and Engineering Research Board (SERB), Department of Science and Technology (DST), Government of India, for supporting the hardware used in this research under the Start-up Research Grant (SRG) scheme.

Declaration on Generative AI

No generative AI tools were used in the preparation of this paper, including in writing, data analysis, figure generation, or any other part of the research process, in accordance with the CEUR Policy.

References

[1] S. Pramanick, N. Sharma, M. S. Akhtar, A. Mukherjee, T. Chakraborty, Detecting harmful memes and counter speech generation: A semi-supervised approach using pre-trained language models, in: Proceedings of the 32nd ACM Conference on Hypertext and Social Media, 2021.

[2] 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: NeurIPS Workshop / Proceedings, 2020. URL: https://arxiv.org/abs/2005.04790.

[3] B. R. Chakravarthi, R. Priyadharshini, R. Ponnusamy, P. Kumaresan, R. Krishnan, et al., Hasoc-dravidiancodemix: Offensive language identification in code-mixed dravidian languages, in: Proceedings of the 4th Workshop on Technologies for Social Media Content Analysis, 2022.

[4] C. Sharma, D. Bhageria, P. Pabreja, R. Goel, M. Singh, et al., Semeval-2020 task 8: Memotion analysis—the visuo-lingual metaphor!, in: Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020), 2020.

[5] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, arXiv preprint arXiv:1703.04009 (2017). URL: https://arxiv.org/abs/1703.04009.

[6] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys (CSUR) 51 (2018) 1–30.

[7] A. Kumar, A. Gupta, M. S. Akhtar, A. Ekbal, T. Chakraborty, Cross-lingual meme classification for low-resource languages, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2021.

[8] 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 (NAACL), 2019.

[9] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, et al., Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020.

[10] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems (NeurIPS), 2012.

[11] L. Gao, P. Zhang, J. Song, H. T. Shen, Complementary multimodal approaches for meme classification, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022.

[12] D. Khattar, J. S. Goud, M. Gupta, V. Varma, Meme and image spam detection using multimodal deep learning, in: Proceedings of the 2019 World Wide Web Conference (WWW), 2019.

[13] A. Das, T. Kumari, N. Sharma, M. S. Akhtar, Hate meme detection in low-resource indic languages: A modular multimodal approach, in: Proceedings of the International Conference on Computational Linguistics (COLING), 2022.

[14] 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.

[15] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[16] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proceedings of the 3rd International Conference on Learning Representations (ICLR), 2015.

[17] S. Khanuja, P. Bansal, P. Mehta, P. Aggarwal, A. Ray, et al., Muril: Multilingual representations for indian languages, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2021.

[18] A. Sarkar, S. Ghosh, A. Ekbal, Emoji-aware preprocessing for low-resource multimodal hate speech detection, in: Proceedings of the European Chapter of the Association for Computational Linguistics (EACL), 2023.

[19] HASOC Shared Task Organizers, Hasoc 2025: Hate speech and offensive content identification in multimodal memes, Forum for Information Retrieval Evaluation (FIRE), 2025.

[20] 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.

[21] A. Kumar, et al., Semeval-2020 task 8: Memotion analysis, in: Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020), 2020. URL: https://aclanthology.org/2020.semeval-1.99.pdf.

[22] S. Mishra, S. Suryavardan, P. Patwa, M. Chakraborty, et al., Memotion 3: Dataset on sentiment and emotion analysis of code-mixed hinglish memes, arXiv preprint arXiv:2309.06517 (2023). URL: https://arxiv.org/abs/2309.06517.

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

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, arXiv preprint arXiv:2103.00020 (2021). URL: https://arxiv.org/abs/2103.00020.

[25] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, Proceedings of Machine Learning Research (2023). URL: https://arxiv.org/abs/2301.12597.

[26] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, et al., Flamingo: a visual language model for few-shot learning, in: NeurIPS, 2022. URL: https://arxiv.org/abs/2204.14198.

[27] 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.

[28] b. 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, CEUR-WS.org, 2025.

[29] A. Bhattacharjee, M. T. Rahman, M. S. I. Azam, S. Rahman, S. Joty, M. S. F. Islamb, M. T. Karim, Banglabert: Language model pretraining and benchmarks for low-resource language understanding, in: Findings of the Association for Computational Linguistics: ACL 2022, 2022, pp. 1054–1069.