Multilingual Hate Speech Classification in Memes Using OCR-Extracted Text and Visual Features

Subham Kumar Rai[1,†], Kunal Goswami[1,†], Gunjan Kumar[2,†], Sangita Singh[2,*,†] and J.P. Singh[2,†]

[1]Sikkim Manipal Institute of Technology, Majitar, Rangpo, East Sikkim, 737136
[2]National Institute of Technology Patna, Ashok Rajpath, 800005

Abstract

The rise of multimodal hate speech in the form of memes and images poses significant challenges to digital safety, particularly in multilingual societies such as South Asian countries like India. Although researchers have widely explored text-based detection, these methods often fall short when hateful content appears in multimodal formats, especially images containing regional scripts. This paper presents a multilingual, multimodal hate speech classification in Regional Languages, including Hindi, Gujarati, Bangla, and Bodo. Using the Hate Speech and Offensive Content (HASOC)-2025 dataset, we extract deep visual features through ResNet-101 and combine them with text features obtained via Optical Character Recognition (OCR). These multimodal features are then used to classify content into two categories: abuse and vulgar, or not abuse and non-vulgar. We balance the class imbalance using ADASYN oversampling, and XGBoost classifiers were trained individually for each language. Our experiments demonstrate that the proposed methodology generalizes effectively across diverse scripts, achieving promising performance despite OCR noise and language-specific challenges. The findings highlight the potential of combining deep image embeddings with classical machine learning for multilingual hate speech classification. Our team, HASOC_2025, achieved some solid results—ranking 15^{th} in Bangla with a score of 0.48746, 13^{th} in Hindi with 0.53602, 15^{th} in Gujarati with 0.34472, and 8^{th} in Bodo with 0.57776.

Keywords

Hate Speech, Memes, OCR, Image, Classification, Regional language

1. Introduction

Hate speech has become one of the most pressing problems in online communication. While abusive and offensive language has been studied extensively in English, less attention has been given to multilingual hate speech detection. This is a particularly critical issue in South Asia, where a large portion of the online population communicates in regional languages such as Hindi [1], Gujarati, Bangla, Marathi [2], Maithili, Assamese, and Bodo. A unique challenge in this context is that hate speech is often not presented as plain text but embedded in images and memes, making detection more complex. Unlike text, which can be processed using NLP (Natural Language Processing)-based models, memes require a combination of optical character recognition (OCR) to extract textual content and computer vision models to capture visual context. Moreover, OCR itself is prone to errors in different scripts:

- Hindi uses the Devanagari script, which OCR systems handle relatively well.
- Bangla uses the Bengali script, which has more complex ligatures, often leading to extraction errors.
- OCR noise arises due to limited availability of pre-trained models for regional scripts.
- Bodo, written in Devanagari, shares some OCR challenges with Hindi but has many fewer resources and datasets.

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

*Corresponding author.

†These authors contributed equally.

subham\protect1_202300274@smit.smu.edu.in (S. K. Rai); kunal\protect1_202300216@smit.smu.edu.in (K. Goswami); gunjank.phd20.cs@nitp.ac.in (G. Kumar); sangitas.ph21.cs@nitp.ac.in (S. Singh); jps@nitp.ac.in (J.P. Singh)

Another major challenge in this task is data imbalance. In real-world hate speech datasets, hateful or abusive content is significantly less frequent than non-hate speech content. Without balancing, classifiers tend to be biased toward majority classes, missing critical hateful expressions.

In response, we propose a multilingual multimodal classification model that can be applied consistently across four languages: Hindi, Gujarati, Bangla, and Bodo. Using pre-trained ResNet-101 embeddings for image features and OCR-extracted text, the proposed model used ADASYN oversampling to balance minority classes and machine learning classifiers for classification. The overall contributions of this paper are as follows:

1. We developed a unified multimodal model that can be effectively applied across four different low-resource Indian languages for hate speech classification.

2. We analyzed the impact of OCR quality on classification performance for each language, highlighting how extraction errors influence downstream results.

3. We demonstrated that classical machine learning, when integrated with image features extracted using deep learning combined with text features, provides robust and competitive performance across multiple languages.

In this work, we proposed a multimodal model that can classify memes into hate and non-hate and generalize across multiple languages and scripts, offering a promising direction for multilingual hate speech detection.

2. Related Work

In this section, we present a concise overview of previous studies focused on the detection of hate speech. Mandl et al. [3] participated in the HASOC track at FIRE 2020, which was focusing on hate speech and offensive content identification in Indo-European languages—including Hindi, German, and English. They used transformer-based models like BERT to classify hate speech for binary and fine-grained classification. However, the performance of their proposed model remained modest and obtained F_1 scores around 0.33 for Hindi fine-grained classification. The obtained results show challenges in low-resource language settings and code-mixed data. Rana and Jha [4] introduced a multimodal deep learning framework combining verbal (text), visual, and acoustic emotion features to detect hate speech in multimedia content. They created the HSDVD video dataset. They also used a Transformer-based model, BERT, to classify hate speech and obtained precision, recall and F_1 scores of 93.00, 92.89 and 92.49, respectively. The main focus was on emotion and multimedia, but they did not explicitly target memes or OCR text embedded in images. Additionally, the dataset used by them was comprised of resource-rich contexts, rather than low-resource languages like Bangla, Gujarati, or Bodo. Modha et al. [5] also participated in FIRE 2022 HASOC to address hate speech and offensive language detection in Hindi-English code-mixed and Marathi. They experimented with multilingual embeddings and transformers for conversational contexts. The limitations of their work were that they focused on text only (code-mixed or monolingual). Visual or multimodal content (e.g., memes) and OCR were not considered in their work. Satapara et al. [6] presented the HASOC 2023 subtrack for hate speech identification in Sinhala and Gujarati, creating benchmark datasets and evaluating multilingual detection approaches. For the competition, several datasets were curated, attracting participation from more than 30 teams that collectively submitted close to 200 runs. The top teams achieved F_1 scores of 0.83 (Sinhala) and 0.84 (Gujarati). The primary limitation of the work was that they relied only on text-based detection, without integrating multimodal features. Although low-resource languages were considered, the absence of visual context limited the scope of the analysis. The HASOC Track at FIRE 2025, having title "Abusive Meme Identification-Shadows Behind the Laughter," [7],[8] shared research task aimed at advancing the development of automated approaches for detecting abusive or offensive content embedded within memes. Organized under the FIRE 2025, this track emphasizes the challenge of multimodal analysis, requiring participants to jointly interpret textual and visual information to effectively identify and classify abusive intent. Ghosh et al. [9] have recognized the critical need to extend efforts to under-resourced languages in order to identify hate speech in low-resource settings. They also analyze inter-system agreement using Cohen's k and Fleiss' k, and investigate item-level difficulty through hardness analysis. Rizwan et al. [10] studied about the growing capabilities of vision-language models (VLMs), which have demonstrated exceptional performance across a wide range of multimodal tasks. they investigated the potential of VLMs for hate meme detection under a zero-shot setting, thereby eliminating the need for task-specific annotated datasets. Through extensive prompt engineering, they systematically evaluate several state-of-the-art VLMs using diverse prompt formulations to classify hateful and harmful memes. They analyze misclassification instances employing a superpixel-based occlusion technique, which provides insights into model behavior and decision-making.

Sahin et al. [11] tackled multimodal hate speech detection, especially in text-embedded images (e.g., memes). Text embedded within images has emerged as a powerful medium for spreading hate speech, extremist ideologies, and propaganda, with notable use during events such as the Russia–Ukraine conflict, where both sides employed such content extensively. Detecting this type of multimodal hate speech is essential to reducing its harmful impact, which has motivated recent research efforts. In the Multimodal Hate Speech Event Detection 2023 shared task, two subtasks were addressed: hate speech detection and target identification. For hate speech detection, researchers adopted multimodal deep learning models enhanced with ensemble strategies and syntactic text features, and they obtained a precision of 84.1, a recall of 89.0, an F_1-score of 84.8, and an accuracy of 84.9. However, for the second subtask, target detection, the multimodal models combined with named entity-based features achieved a precision of 76.36, recall of 76.37, F_1-score of 76.34, and accuracy of 79.34. The results indicated that the proposed models consistently surpassed text-only, image-only, and multimodal baselines, securing the top position in both subtasks.

Mandl et al. [12] worked on the shared task, which includes three subtasks, addressing both binary and detailed category classification. In total, 321 experimental runs were submitted in these subtasks. Most of the participating systems relied on LSTM-based models that utilized word embeddings as features. The leading approaches achieved Macro-F_1 scores of 0.78 for English, 0.81 for Hindi, and 0.61 for German, indicating clear differences in performance between languages.

Several prior studies have explored hate speech classification using multilingual and cross-lingual approaches in Indic low-resource languages[13, 14, 15]. Ghosh et al. [16] investigated the effectiveness of transformer architectures—including BERT, RoBERTa, ALBERT, and DistilBERT—on existing Indian hate speech datasets, namely HASOC-Hindi (2019), HASOC-Marathi (2021), and Bengali Hate Speech (BenHateSpeech), focusing on binary classification. They found that traditional deep learning approaches often struggle to detect hateful expressions when offensive terms are embedded within complex or nuanced phrasing. In contrast, transformer models are capable of understanding contextual semantics, allowing them to more effectively identify hate speech even in subtle linguistic settings. Their proposed work compares multilingual transformer models such as MuRILBERT and XLM-RoBERTa with monolingual models, NeuralSpaceBERTHi (Hindi), MahaBERT (Marathi), and BanglaBERT (Bengali). Experimental results indicate that the monolingual MahaBERT achieves superior performance on the HASOC-Marathi dataset, while MuRILBERT outperforms other models on HASOC-Hindi and BenHateSpeech. Additionally, cross-lingual evaluations between Hindi and Marathi models reveal a mix of consistent and divergent outcomes, highlighting important linguistic and contextual variations across Indian languages.

Ranasinghe et al. [17] address this gap, HASOC 2023 organized a series of shared tasks aimed at detecting offensive and hateful content in such languages, which focused on hate speech detection in several Indo-Aryan languages—namely Assamese, Bengali, Gujarati, and Sinhala—as well as in Bodo, a Sino-Tibetan language with scarce linguistic resources. The shared task facilitated the development of multiple curated datasets for these languages, encouraging multilingual and cross-lingual experimentation. In total, approximately 200 runs were submitted by over 30 participating teams, whose results and comparative analyses are summarized and discussed in this report.

Kashif et al. [18] have proposed an ensemble-based framework to classify text-embedded images into two categories: Hate Speech and No Hate Speech. The approach integrates advanced models such as InceptionV3, BERT, and XLNet, leveraging their combined strengths for multimodal analysis. Experimental evaluations demonstrated encouraging outcomes, achieving an accuracy of 75.21 and an F_1-score of 74.96, thereby highlighting the effectiveness of ensemble learning in detecting and classifying hateful content within social media posts. The proposed work was limited to English-language memes. Ganguly et al. [19] participated in the Multimodal Hate Speech Event Detection shared task. In the Shared Task on Multimodal Hate Speech Event Detection at CASE 2024 (EACL 2024), they introduced the MasonPerplexity system to tackle this issue. The competition consists of two subtasks: Subtask B, which seeks to locate targets in text-embedded images during political events, and Subtask A, which focuses on identifying hate speech. They used an ensemble of XLM-RoBERTa-base, BERTweet-large, and BERT-base for Subtask B and an XLM-RoBERTa-large model for Subtask A. Their suggested approach placed third in both categories with F_1-scores of 0.8347 for Subtask A and 0.6741 for Subtask B.

Chhabra and Vishwakarma [20] introduced a scalable framework for multimodal hate content identification called as Scalable Transformer-based Multilevel Attention (STMA). This model includes three major components: A combined attention-driven deep learning module, a vision attention encoder, and a caption attention encoder are all designed to handle multimodal information using distinct attention processes. This architecture's effectiveness was evaluated on three benchmark datasets: Hateful Memes, MultiOff, and MMHS150K, utilizing a variety of measures. STMA consistently beat baseline models on all datasets, suggesting its robustness and potential for detecting multimodal hate speech. Sultana et al. [21] investigate hate speech detection on Twitter, going beyond simply identifying damaging content to include recognizing targeted groups and the intensity of aggression expressed.

To address this, a fusion-based deep learning system is presented, which combines BERT and fastText embeddings to capture both contextual meaning and subword-level characteristics. Experiments conducted on the SemEval 2019 Task 5 dataset show that the proposed model surpasses prior methods in Subtask B (Target and Aggression Classification) with a 67% F_1-score, while also delivering competitive results in Subtask A (Hate Speech against immigrants and women), and obtained a F_1-score of 65%. The findings of their research highlight the strength of embedding fusion and provide meaningful contributions toward more comprehensive hate speech detection by considering both the audience and the aggressiveness of language.

A composite feature fusion framework, the Multi-Head Attention with LSTM (MHA-LSTM) model, was presented by Kalaivani et al.[22]. It blends attention-driven deep learning architecture with local-global feature extraction. In that approach, a 1D-CNN was used to capture both short-range and long-range textual dependencies, and the extracted features were merged into a unified representation. The representation was then processed through the MHA-LSTM module, which uses attention mechanisms to highlight the most informative aspects of the input for hate speech detection. The method was tested on the HASOC 2021 Dravidian dataset containing Tamil code-mixed social media comments. Experimental analysis revealed strong performance, with the model achieving 95.6% accuracy and surpassing several state-of-the-art systems across multiple evaluation metrics, including precision, recall, specificity, and F_1-score. The findings confirm that integrating local global feature fusion with MHA-LSTM substantially enhances offensive content detection in Tamil code-mixed text, especially under a 90/10 train–test split.

A multimodal hate speech detection framework using a late fusion method that combines Wav2Vec 2.0 for speech representation with Muril for text analysis was presented by Selvamurugan et al.[23]. The DravidianLangTech@NAACL 2025 dataset, which comprises text and speech data for Telugu, Tamil, and Malayalam and is divided into six groups, was used to assess the method: Religious hate, political hate, gender hate, personal defamation, and non-hatred. By combining data augmentation and class weighting strategies, they addressed class imbalance. The relevance of multimodal tactics for enhancing detection in low-resource language environments is shown by their experimental results, which show that the late fusion strategy successfully identifies hate speech patterns that would be missed when utilizing a single modality.

3. Methodology

The proposed framework follows a multimodal approach that integrates textual and visual context for effective hate speech classification in low-resource regional languages. The methodology consists of a detailed description of the used dataset 3.1 and the proposed method which consists of five main stages: preprocessing 3.2, feature extraction 3.3, feature fusion with class balancing 3.4, and classification, as shown in Figure1.

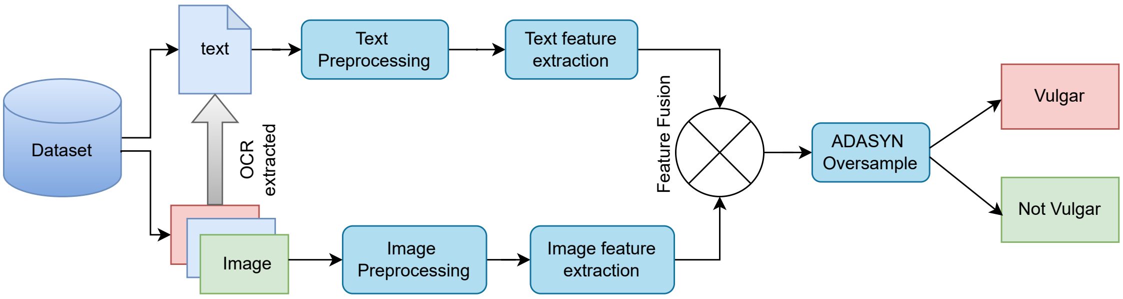

Figure 1: The proposed model diagram used for classifying memes

.

3.1. Dataset discription

The dataset is taken from the HASOC-2025, organized by FIRE. The dataset is available in four Indo-Aryan (low-resourced) languages, such as Gujarati, Hindi, Bangla and Bodo. The provided dataset is a multimodal dataset which consists of memes containing both images and embedded text. In the Gujarati, Bangla and Bodo languages, the textual content was on the image, while for the Hindi language, the organizer provided text in separate CSV file having Image ID, text and their corresponding labels. However, for the other three languages, they provided only image ID and their corresponding labels. Each datasets is label multiple categories like Sentiment, Sarcasm, Vulgar, Abuse, and Target.

3.2. Preprocessing

The dataset consists of memes containing both images and embedded text across four languages: Hindi, Bangla, Gujarati, and Bodo. Since textual content is often embedded within images, Optical Character Recognition (OCR) was applied to extract the textual components. For languages written in Devanagari (Hindi and Bodo), OCR produced relatively accurate results, whereas Bangla and Gujarati posed higher challenges due to complex ligatures and limited OCR resources. The extracted text was further cleaned through the removal of stop-words, emojis, punctuation and tokenization. For the image preprocessing, we resize all the images to a size of 224×224.

3.3. Feature Extraction

After preprocessing, we extracted the features from both modalities, such as text and images. For the textual modality, OCR is applied to extract language-specific text from memes. After that, the text features were extracted using (Term Frequency-Inverse Document Frequency) TF-IDF, BERT, GloVe [24], and FastText. It is found that the TF-IDF is performing best because it finds the relationship of an occurrence of a particular word in the data sample with respect to the whole dataset. The visual modality of memes is processed using a pre-trained ResNet-101 [25] convolutional neural network, which generated 2048-dimensional embeddings for each image. The generated high-dimensional embeddings help to capture semantic and contextual details from the images. These embeddings form a rich feature space, enabling the model to detect subtle visual cues often associated with hateful content. These embeddings captured complementary semantic and contextual information from both modalities.

Table 1
Performance Comparison Across Languages

Language	Text-only (Accuracy %)	Image-only (Accuracy %)	Fusion (Accuracy %)	Fusion (Macro-F_1)
Hindi	75	73	85	0.83
Bangla	70	68	82	0.80
Gujrati	65	62	75	0.72
Bodo	60	58	72	0.70

3.4. Feature Fusion and Class Balancing

We concatenated the features extracted from the text and the image to identify the hate speech more accurately than the individual ones. The text and image features were concatenated to make a single vector representation. After concatenating the feature vector, it is passed to the machine learning models for classification.

After that, we found that provided multiclass hate speech datasets are typically imbalanced, with hateful categories underrepresented relative to neutral content. To address this, we incorporate the Adaptive Synthetic Sampling (ADASYN) oversampling method [26], which generates synthetic examples of minority classes. This ensures that the classifier does not become biased toward majority classes, improving the recall for hateful categories. This ensured that the classifier received a balanced distribution of training examples.

3.5. Classification

We divided each dataset into the ratio of training (70%), validation (15%), and testing (15%), and used stratified sampling to maintain class balance. We used early stopping and learning rate schedulers to avoid overfitting and hyperparameter tuning was done for batch size, learning rate, and optimizer (Adam/SGD).

Text-Only Approach: In the first stage, we utilized textual data to perform multi-class hate speech classification. We used eight traditional machine learning algorithms-(1) Logistic Regression, (2) Support Vector Machine, (3) Naïve Bayes, (4) K-Nearest Neighbors (KNN), (5) Decision Tree, (6) Random Forest, (7) Gradient Boosting, and (8) XGBoost-as well as two deep learning models, namely Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) [27] networks. Now onwards, we used only XGBoost for further experiments.

Image-Only Approach: In the next phase, we relied solely on image features for hate speech classification using ResNet and CNN models. However, the performance was comparatively weaker, with the accuracy dropping by approximately 2% for each language, as summarized in Table 1.

Fusion-based Approach: Once features from both modalities are fused, the fused vector (ResNet embeddings + TF-IDF features) is passed to the machine learning classifiers. In particular, XGBoost performed well because of its scalability and ability to handle heterogeneous input features. XGBoost is advantageous in this setting because it can learn complex non-linear decision boundaries without requiring extensive computational resources compared to end-to-end deep learning classifiers.

4. Results

The proposed model was validated with four low-resource languages, Bangla, Gujarati, Hindi, and Bodo, datasets using three model configurations: Text-only, Image-only, and Multimodal Fusion. This section presents the outcomes of these experiments, comparing performance across languages and modalities.

For each language, we report classification accuracy, precision, recall, and F_1-scores, to highlight model performance on minority classes. The results demonstrate the effectiveness of combining ResNet-101 image embeddings with TF–IDF textual features and an XGBoost classifier balanced using ADASYN.

Table 2
Performance metrics for Hindi, Bangla, Gujrati and Bodo dataset with proposed model (ResNet-101 + OCR + XGBoost)

Language	Accuracy	Precision	Recall	F1-score
Hindi	0.86	0.84	0.83	0.83
Bangla	0.81	0.79	0.77	0.78
Gujrati	0.79	0.76	0.75	0.75
Bodo	0.74	0.72	0.70	0.71

For Hindi memes, the model achieved high accuracy with balanced results across hate and non-hate categories. However, the F_1-score for hate speech was slightly lower, mainly due to OCR inaccuracies in Devanagari text extraction. Despite this, the multimodal approach effectively captured relevant features, with the confusion matrix showing relatively few false negatives. In Bangla, performance declined because of the script's complexity, which introduced OCR errors that propagated to the TF-IDF representation, reducing recall for the hate class. Although overall accuracy was competitive, the F_1-score for hateful content was modest, underscoring the challenges posed by ligatures and script intricacies and the need for better OCR tools for Indic scripts. The Gujarati dataset posed additional difficulties due to scarce pretrained OCR resources. While the classifier performed reasonably, the hate class showed higher misclassification rates, leading to a moderate overall F_1-score. Still, the integration of visual and textual features outperformed text-only baselines. For Bodo, which also uses Devanagari but remains low-resource, the model achieved lower accuracy and recall, with a bias toward the majority non-hate class. This outcome was largely due to limited annotated data and OCR errors in less common script variants, restricting the model's ability to generalize effectively.

Overall, Hindi achieved the best results, followed by Gujarati and Bangla, with Bodo showing the weakest performance due to resource scarcity as shown in Table 2. The multimodal fusion pipeline demonstrated robustness by improving over text-only approaches, but OCR errors and dataset imbalance continue to limit performance. These findings suggest that resource availability and OCR quality are the strongest predictors of performance across Indic languages.

The results obtained by the teams in FIRE: HASOC-2025 are shown in the Tables 3 and 4. The results show that FiRC-NLP consistently dominated in most languages, securing the highest scores of 0.62755 in Bangla, 0.65706 in Hindi, and 0.67501 in Gujarati, while NLPFusion led in Bodo with a score of 0.63128. At the other extreme, the lowest scores were reported by DeepSemantics in Bangla (0.48211), NITA_ICFAI in Hindi (0.34037), and our own secondary submission in Gujarati (0.34472). For Bodo, the baseline system HASOC2025-meme obtained the lowest score of 0.39221. Our submissions achieved 0.53185 (12th) in Bangla, 0.54181 (11th) in Hindi, 0.49293 (12th) and 0.34472 (14th) in Gujarati, and 0.57776 (8th) and 0.39221 (15th) in Bodo. These results indicate that although our models did not reach the top ranks, they performed competitively in low-resource settings, particularly in Bodo, where our system, with a score of 0.57776, ranked within the top ten.

5. Discussion

The proposed hybrid ML–DL pipeline offers several advantages by using deep learning only for feature extraction and relying on XGBoost for classification. This design captures rich semantic and visual representations without the computational overhead of full end-to-end training, making the framework efficient and well-suited for low-resource languages. While deep learning extracted embedding vectors ensure robust feature extraction for image data, XGBoost provides lightweight yet effective classification, striking a balance between accuracy. This makes the approach particularly valuable for languages such as Bodo and Gujarati, which face resource constraints, while still maintaining strong generalization for Bangla and Hindi. The experimental evaluation confirms the effectiveness of combining deep learning-based embeddings with machine learning-based classifiers for multilingual multimodal hate speech detection. The integration of ResNet-101 image embeddings with TF–IDF features from OCR-

Table 3
The rank achieved by our team in FIRE: HASOC-2025 for Bangla and Hindi languages

Language	Rank	Team_name	Score
Bangla	1	FiRC-NLP	0.62755
	2	CSIS BITS Pilani	0.61820
	3	Golden Ratio	0.61538
	4	SCaLAR	0.61034
	5	NLPFusion	0.60834
	6	KK_NLP_AI_IIIT_Ranchi	0.60595
	7	Abu Taher	0.57563
	8	IIT Dhanbad	0.57209
	9	DeepMeme	0.56203
	10	CSE_SVNIT	0.55476
	11	MUCS	0.53785
	12	**HASOC2025_meme(Baseline)**	**0.53185**
	13	kongqiang wang	0.52528
	14	IReL	0.50818
Our	15	**HASOC_2025**	**0.48746**
	16	Charmathi Rajkumar	0.48331
	17	Team DeepSemantics	0.48211
Hindi	1	FiRC-NLP	0.65706
	2	NLPFusion	0.62398
	3	KK_NLP_AI_IIIT_Ranchi	0.59597
	4	Golden Ratio	0.59097
	5	SCaLAR	0.58468
	6	IIT Dhanbad	0.57417
	7	CSE_SVNIT	0.57198
	8	Peng Zhang	0.57094
	9	CSIS BITS Pilani	0.56788
	10	Charmathi Rajkumar	0.56769
	11	Team DeepSemantics	0.54989
	12	**HASOC2025-meme(Baseline)**	**0.54181**
Our	13	**HASOC_2025**	**0.53602**
	14	FAST	0.52881
	15	MUCS	0.52497
	16	kongqiang wang	0.51985
	17	IReL	0.46302
	18	NITA_ICFAI	0.34037

processed text consistently improves performance across Hindi, Bangla, Gujarati, and Bodo. The use of ADASYN further alleviates class imbalance, boosting recall for minority hate speech categories. Results (Table 1) show that multimodal fusion consistently outperforms unimodal setups, achieving higher accuracy and macro-F_1 across all languages. For instance, fusion accuracy reached 85% for Hindi and 82% for Bangla, compared to lower scores in text-only and image-only baselines. Similarly, the detailed metrics (Table 2 show robust precision, recall, and F_1-scores, with Hindi achieving 0.83 macro-F_1, followed by Bangla (0.78), Gujarati (0.75), and Bodo (0.71). Despite these strengths, the system faces three key limitations. First, OCR introduces significant noise, particularly for Bangla and Gujarati, where complex ligatures and underdeveloped OCR resources often yield incomplete or inaccurate text. These errors propagate downstream, affecting classifier reliability. Second, the dependency on TF–IDF restricts textual representation to surface-level frequency patterns, limiting the ability to capture deeper contextual and semantic relationships, especially in multilingual and code-mixed data. Third, while the hybrid design balances performance and efficiency, it does not exploit advanced end-to-end multimodal transformer architectures, which could better capture cross-modal interactions. Furthermore, severe data scarcity, especially in Bodo, restricts generalizability and highlights the urgent need for larger annotated multimodal corpora. Overall, the findings demonstrate that the hybrid pipeline offers a

Table 4
The rank achieved by our team in FIRE: HASOC-2025 for Gujarati and Bodo languages

Language	Rank	Team_name	Score
Gujarati	1	FiRC-NLP	0.67501
	2	NLPFusion	0.63436
	3	MUCS	0.61848
	4	SCaLAR	0.61715
	5	KK_NLP_AI_IIIT_Ranchi	0.61409
	6	CSIS BITS Pilani	0.60018
	7	IReL	0.59196
	8	IIT Dhanbad	0.58879
	9	CSE_SVNIT	0.58221
	10	kongqiang wang	0.56253
	11	Peng Zhang	0.55522
	12	Charmathi Rajkumar	0.54678
	13	**HASOC2025-meme(Baseline)**	**0.49293**
	14	Team DeepSemantics	0.42035
Our	15	**HASOC_2025**	**0.34472**
Bodo	1	NLPFusion	0.63128
	2	FiRC-NLP	0.62217
	3	CNLP-UPES (Pankaj Dadure)	0.60921
	4	SCaLAR	0.60393
	5	CSIS BITS Pilani	0.59969
	6	CSE_SVNIT	0.58730
	7	IIT Dhanbad	0.58186
Our	8	**HASOC_2025**	**0.57776**
	9	KK_NLP_AI_IIIT_Ranchi	0.57184
	10	Golden Ratio	0.56202
	11	Team DeepSemantics	0.5604
	12	MUCS	0.55215
	13	Charmathi Rajkumar	0.54566
	14	IReL	0.50111
	15	**HASOC2025-meme(Baseline)**	**0.39221**

practical balance between efficiency and accuracy for low-resource languages, but addressing OCR errors, enhancing text modelling, and expanding annotated datasets remain open challenges for future work.

5.1. Conclusion

This study combined OCR-based text extraction, TF-IDF features, ResNet-101 image embeddings, ADASYN oversampling, and XGBoost classification to present a multilingual multimodal framework for hate speech detection in four South Asian languages: Hindi, Bangla, Gujarati, and Bodo. The results show that multimodal hate speech can be efficiently addressed in low-resource environments by lightweight machine learning classifiers enhanced with deep visual embeddings. The framework is especially well-suited for languages with few annotated resources because it achieves a realistic balance between accuracy and computational performance. A number of improvements can be investigated for further investigation. Textual feature representations of OCR outputs could be significantly enhanced by using transformer-based contextual embeddings like XLM-R or IndicBERT, particularly for languages with complicated morphology or code. By capturing intricate cross-modal dependencies, multimodal transformers that simultaneously describe textual and visual modalities may improve performance even more. Another encouraging approach is to enhance OCR pipelines for Indic scripts by creating synthetic data or fine-tuning them specifically for each script. Building more stable and generalizable systems would also require growing annotated multimodal datasets, especially for low-resource languages like Bodo. In summary, while the proposed framework provides an efficient solution for multilingual multimodal hate speech detection, advancements in OCR, multimodal transformer architectures, and dataset availability will be key to broadening its applicability and achieving stronger performance across diverse linguistic settings.

Declaration on Generative AI

This paper includes no content generated by artificial intelligence tools beyond language editing and formatting assistance. All intellectual contributions, including the conception, analysis, and interpretation of results, were made by the authors.

References

[1] G. Kumar, J. P. Singh, Hate speech and offensive content identification in english and indo-aryan languages using machine learning models, in: K. Ghosh, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, Kolkata, India, December 9-13, 2022, volume 3395 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 542–551. URL: https://ceur-ws.org/Vol-3395/T7-7.pdf.

[2] A. Kumari, A. Garge, P. Raj, G. Kumar, J. P. Singh, M. Alryalat, Classification of offensive tweet in marathi language using machine learning models, in: K. Dasgupta, S. Mukhopadhyay, J. K. Mandal, P. Dutta (Eds.), Computational Intelligence in Communications and Business Analytics - 5th International Conference, CICBA 2023, Kalyani, India, January 27-28, 2023, Revised Selected Papers, Part I, volume 1955 of *Communications in Computer and Information Science*, Springer, 2023, pp. 261–273. URL: https://doi.org/10.1007/978-3-031-48876-4_20. doi:10.1007/978-3-031-48876-4_20.

[3] 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: D. Ganguly, S. Gangopadhyay, M. Mitra, P. Majumder (Eds.), FIRE 2021: Forum for Information Retrieval Evaluation, Virtual Event, India, December 13 - 17, 2021, ACM, 2021, pp. 1–3. URL: https://doi.org/10.1145/3503162.3503176. doi:10.1145/3503162.3503176.

[4] A. Rana, S. Jha, Emotion based hate speech detection using multimodal learning, CoRR abs/2202.06218 (2022). URL: https://arxiv.org/abs/2202.06218. arXiv:2202.06218.

[5] S. Satapara, P. Majumder, T. Mandl, S. Modha, H. Madhu, T. Ranasinghe, M. Zampieri, K. North, D. Premasiri, Overview of the HASOC subtrack at FIRE 2022: Hate speech and offensive content identification in english and indo-aryan languages, in: D. Ganguly, S. Gangopadhyay, M. Mitra, P. Majumder (Eds.), Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE 2022, Kolkata, India, December 9-13, 2022, ACM, 2022, pp. 4–7. URL: https://doi.org/10.1145/3574318.3574326. doi:10.1145/3574318.3574326.

[6] T. Ranasinghe, K. Ghosh, A. S. Pal, A. Senapati, A. E. Dmonte, M. Zampieri, S. Modha, S. Satapara, Overview of the HASOC subtracks at FIRE 2023: Hate speech and offensive content identification in assamese, bengali, bodo, gujarati and sinhala, in: D. Ganguly, S. Majumdar, B. Mitra, P. Gupta, S. Gangopadhyay, P. Majumder (Eds.), Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE 2023, Panjim, India, December 15-18, 2023, ACM, 2023, pp. 13–15. URL: https://doi.org/10.1145/3632754.3633278. doi:10.1145/3632754.3633278.

[7] K. Ghosh, M. Das, S. Patel, N. Bhandary, A. Das, A. Mukherjee, S. Modha, D. Ganguly, U. Garain, S. Jaki, et al., Overview of the hasoc track at fire 2025: Abusive meme identification—shadows behind the laughter, in: Proceedings of the 17th annual meeting of the Forum for Information Retrieval Evaluation, 2025, pp. 28–31.

[8] 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.

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

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

[11] U. Sahin, I. E. Kucukkaya, O. Ozcelik, C. Toraman, ARC-NLP at multimodal hate speech event detection 2023: Multimodal methods boosted by ensemble learning, syntactical and entity features, CoRR abs/2307.13829 (2023). URL: https://doi.org/10.48550/arXiv.2307.13829. doi:10.48550/ARXIV.2307.13829. arXiv:2307.13829.

[12] 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.

[13] 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.

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

[15] 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.

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

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

[18] M. Kashif, M. Zohair, S. Ali, Lexical squad@multimodal hate speech event detection 2023: Multimodal hate speech detection using fused ensemble approach, CoRR abs/2309.13354 (2023). URL: https://doi.org/10.48550/arXiv.2309.13354. doi:10.48550/ARXIV.2309.13354. arXiv:2309.13354.

[19] A. Ganguly, A. N. B. Emran, S. S. C. Puspo, M. N. Raihan, D. Goswami, M. Zampieri, Masonperplexity at multimodal hate speech event detection 2024: Hate speech and target detection using transformer ensembles, in: A. Hürriyetoglu, H. Tanev, S. Thapa, G. Uludogan (Eds.), Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text, CASE 2024, St. Julians, Malta, March 22, 2024, Association for Computational Linguistics, 2024, pp. 125–131. URL: https://aclanthology.org/2024.case-1.17.

[20] A. Chhabra, D. K. Vishwakarma, MHS-STMA: multimodal hate speech detection via scalable transformer-based multilevel attention framework, CoRR abs/2409.05136 (2024). URL: https:

//doi.org/10.48550/arXiv.2409.05136. doi:10.48550/ARXIV.2409.05136. arXiv:2409.05136.

[21] S. J. Sultana, M. Nesarul Hoque, A. N. Chy, M. Hanif Seddiqui, Enhanced hate speech detection through mean-pooling in embedding fusion, in: 2024 27th International Conference on Computer and Information Technology (ICCIT), 2024, pp. 1833–1838. doi:10.1109/ICCIT64611.2024.11021822.

[22] A. Kalaivani, D. Thenmozhi, Composite feature fusion for improved offensive language detection in tamil social media using mha-lstm, International Journal of Machine Learning and Cybernetics (2025) 1–17.

[23] A. Selvamurugan, DravLingua@DravidianLangTech 2025: Multimodal hate speech detection in Dravidian languages using late fusion of muril and Wav2Vec models, in: B. R. Chakravarthi, R. Priyadharshini, A. K. Madasamy, S. Thavareesan, E. Sherly, S. Rajiakodi, B. Palani, M. Subramanian, S. Cn, D. Chinnappa (Eds.), Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico, 2025, pp. 694–699. URL: https://aclanthology.org/2025.dravidianlangtech-1.118/. doi:10.18653/v1/2025.dravidianlangtech-1.118.

[24] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, ACL, 2014, pp. 1532–1543. URL: https://doi.org/10.3115/v1/d14-1162. doi:10.3115/V1/D14-1162.

[25] U. Panigrahi, P. K. Sahoo, M. K. Panda, G. Panda, A resnet-101 deep learning framework induced transfer learning strategy for moving object detection, Image and Vision Computing 146 (2024) 105021. URL: https://www.sciencedirect.com/science/article/pii/S0262885624001252. doi:https://doi.org/10.1016/j.imavis.2024.105021.

[26] I. Dey, V. Pratap, A comparative study of smote, borderline-smote, and adasyn oversampling techniques using different classifiers, in: 2023 3rd International Conference on Smart Data Intelligence (ICSMDI), 2023, pp. 294–302. doi:10.1109/ICSMDI57622.2023.00060.

[27] Y. Yu, X. Si, C. Hu, J. Zhang, A review of recurrent neural networks: Lstm cells and network architectures, Neural Computation 31 (2019) 1235–1270. URL: https://doi.org/10.1162/neco_a_01199. doi:10.1162/neco_a_01199. arXiv:https://direct.mit.edu/neco/article-pdf/31/7/1235/1053200/neco$_{a0}1199.pdf$.