Bodo Meme Classification using Multimodal Fusion and Oversampling in HASOC-2025

Pankaj Dadure[1,*,†], Sourav Ghosh[2,†]

[1]UPES Dehradun, Uttrakhand, India

[2]Department of Computer Science and Engineering- AI, Brainware University, West Bengal, India

Abstract

The rapid growth of multimodal content on social media has amplified the need for effective systems that can detect sentiment, sarcasm, vulgarity, abuse, and targeted communities in memes. This study focuses on the under-explored Bodo language, a low-resource setting where the scarcity of annotated data presents significant challenges for meme classification. We propose a multimodal framework that integrates BERT-based OCR text embeddings and ResNet-derived image features through an early-fusion strategy, enabling multi-task predictions across five classification subtasks. To address class imbalance, we incorporate RandomOverSampler, thereby enhancing the representational balance of minority categories. Experimental evaluations on the HASOC-Meme 2025 dataset demonstrate that our system (CNLP-UPES) achieved an F1 score of 0.60921, ranking third among fifteen competing teams. Results highlight the effectiveness of multimodal fusion, with performance gains of 45–55% over unimodal baselines, and establish the robustness of our approach in low-resource multimodal scenarios. Furthermore, comparative analysis reveals that while the top-ranked team achieved a slightly higher score, the marginal difference suggests that improvements primarily arise from pretraining strategies and hyperparameter optimization rather than architectural innovations. This work marks one of the first systematic efforts toward multimodal meme classification in the Bodo language, contributing valuable insights into multimodal learning for low-resource languages.

Keywords

Multimodal Fusion, Oversampling, Bodo Language, Meme Classification, HASOC, Low-Resource NLP

1. Introduction

The rise of social networking platforms such as Twitter, Facebook, and Instagram has transformed the way people communicate, share ideas, and engage in public discourse [1]. The accessibility and interactive nature of these platforms have generated vast amounts of user-generated content, fueling vibrant discussions but also giving rise to harmful and offensive material such as hate speech. Hate speech is defined as any form of communication that attacks or demeans individuals or groups based on attributes such as race, religion, gender, or political affiliation [2]. Such content can erode the quality of dialogue, foster polarization, and undermine democratic ideals. As Habermas [3] argued, public opinion in a democracy depends on rational and inclusive discourse, and the spread of toxic, derogatory, and divisive material poses a direct threat to this foundation. There is a strong need to identify and detect hate speech because its spread harms open discussion and social harmony, and many researchers are already working on creating reliable methods to deal with this problem.

The challenge of detecting harmful content has become more difficult with the way communication has evolved online. Earlier, most hate speech detection methods focused only on text [4], but now malicious users often turn to multimodal formats like memes to escape moderation. Memes usually mix short pieces of text with images that carry cultural meaning, making it possible to hide harmful or coded messages in ways that are hard to catch from text or images alone. Studies shown in [5, 6] highlight that more than 40% of hateful memes look harmless if we look at just the text without the

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

[*]Corresponding author.

[†]These authors contributed equally.

krdadure@gmail.com (P. Dadure); souravg256@gmail.com (S. Ghosh)

0000-0002-8003-9671 (P. Dadure); 0000-0002-6768-0843 (S. Ghosh)

image, or the image without the text, which shows why single-mode detection is not enough. For example, a picture of a political leader combined with ironic praise can actually be harsh criticism, while an ordinary stock photo with a few suggestive words can send a xenophobic message. Because of this, multimodal memes make automated moderation far more challenging. Multimodal hate speech detection [7] addresses this gap by jointly analyzing the semantics of text and the contextual, cultural, and symbolic cues embedded in images. Such systems are better equipped to detect implicit abuse, sarcasm, vulgarity, and targeted harassment, all of which may be subtle yet socially damaging. This approach not only improves detection accuracy but also helps maintain a balance between safety and freedom of expression an essential consideration for open digital societies.

In this landscape, the Hate Speech and Offensive Content Identification (HASOC) 2025 [8, 9] shared task provides a rigorous benchmark for developing and evaluating multimodal moderation systems, requiring participants to solve five interconnected classification tasks:

1. **Sentiment Detection:** Positive, Neutral, or Negative tone.
2. **Sarcasm Detection:** Sarcastic or Non-sarcastic intent.
3. **Vulgarity Detection:** Presence or absence of explicit expressions.
4. **Abuse Detection:** Abusive or Non-abusive content.
5. **Target Community Identification:** Identifying if the meme targets categories such as Gender, Religion, Individual, Political, National Origin, Social Sub-groups, Others, or None.

The HASOC 2025 dataset includes memes annotated across all five dimensions, reflecting real-world complexity in online communication. Solving this challenge requires robust multimodal fusion methods capable of integrating textual features often extracted via Optical Character Recognition (OCR) with visual features derived from image analysis. Crucially, the system must model the interplay between modalities, as meaning often emerges only in their combination. HASOC organized the shared task for four languages such as Bangla, Hindi, Gujarati, and Bodo, while the work presented in this paper focuses exclusively on the Bodo language.

This paper introduces a deep learning framework tailored to the HASOC 2025 challenge. We propose a multimodal classification framework that integrates textual and visual features using an early fusion strategy, where textual representations are derived from a pre-trained BERT model and visual features are extracted using a ResNet architecture. We address the challenge of class imbalance by employing oversampling techniques to ensure equitable representation of minority classes, thereby improving model generalization and robustness. We implement an end-to-end training pipeline, including optimized hyperparameter scheduling and weighted loss functions, to effectively learn from both textual and visual modalities. The rest of this paper is organized as follows. Section 2 reviews related work on hate speech and offensive content detection with a particular focus on multimodal approaches. Section 3 describes the dataset used in this study, highlighting its characteristics and annotation details. Section 4 presents the proposed system architecture, including preprocessing, imbalance handling, and the multimodal fusion framework. Section 5 discusses the experimental setup and evaluation metrics, followed by Section 6 concludes the paper and outlines directions for future work.

2. Related work

Meme analysis [10] has emerged as a challenging task in natural language processing (NLP) and computer vision due to its inherently multimodal nature, combining both visual and textual cues. Early studies in this area primarily focused on unimodal approaches [11], where sentiment classification was attempted either through textual content extracted via OCR or through visual features alone. While these methods provided initial insights, they often failed to capture the nuanced interplay between modalities, resulting in limited performance across diverse datasets and languages. For instance, Keswani et al. [12] tackle meme sentiment classification (SemEval-2020 Task 8, Subtask A) by comparing unimodal and bimodal approaches. They evaluated methods ranging from Naïve Bayes and FFNN with Word2Vec (text-only) to Transformer-based baselines and simple visual pipelines. Notably, the text-only FFNN using Word2Vec embeddings outperformed all other variants including multimodal methods achieving a 63% relative improvement over the baseline macro-F1 score. Hazman et al. [13] propose a supervised intermediate training strategy for multimodal meme sentiment classification by incorporating unimodal sentiment-labeled data (image-only and text-only) as intermediate tasks leveraging the STILT (Supplementary Training on Intermediate Labeled-data Tasks) paradigm. Their method demonstrates a statistically significant performance gain, and importantly, shows that the model's effectiveness remains intact even when the labeled meme training set is reduced by 40%. Behera et al. [14] found that text-only models and image-only models both underperformed compared to a multimodal architecture so they employs attention mechanisms over both modalities. Their proposed multimodal framework achieved a macro-F1 score of 34.23% and accuracy of 50.02%, which corresponds to +6.8% and +7.9% absolute improvements over the best-performing text-only and image-only models, respectively. Unimodal models are limited by their inability to capture cross-modal dependencies. Text-only models suffer from noisy OCR and miss visual cues, while image-only models fail to interpret sarcasm, idioms, or context expressed through text, leading to reduced generalization and accuracy.

To overcome these limitations, researchers shifted toward multimodal models, leveraging both image and text features for improved meme understanding. Several works have explored fusion strategies [15][16][17], such as early fusion [18], where textual and visual embeddings are combined at the input stage, and late fusion [18], where modality-specific predictions are integrated at the decision level. These methods demonstrated that multimodality leads to more robust representations, though their performance often varied across languages and cultural contexts. For instance, Guo et al. [19] tackles meme emotion classification (SemEval-2020 Task 8) using a simple concatenation-based multimodal fusion of text embeddings (via BERT) [20] and image features (via DenseNet). Their ablation showed that DenseNet consistently outperformed ResNet in image-only configurations and, interestingly, that integrating text through BERT often did not yield further improvements. This suggests that in the context of meme sentiment, visual features may carry more discriminative signal than textual ones. Sharma et al. [15] tackled the problem of meme emotion detection, where prior multimodal models struggled with affective grounding and generalization. They introduced MOOD, a dataset annotated with six basic emotions, and proposed ALFRED, a multimodal fusion framework that integrates emotion-enriched visual cues with text through a gating mechanism. ALFRED outperformed baselines by 4.94% F1, showed strong results on Memotion, and generalized effectively to HarMeme and Dank Memes, Chauhan et al. [21] introduced a deep attentive multi-task framework for meme understanding, where five tasks—humour, sarcasm, offensiveness, motivation, and sentiment are learned jointly. The model employs two novel attention modules: the Inter-task Relationship Module (iTRM) to capture dependencies across tasks, and the Inter-class Relationship Module (iCRM) to model class-level correlations. This multi-task fusion significantly outperformed single-task models and SemEval-2020 baselines on the Memotion dataset. Alzu'bi et al. [22] tackled offensive meme detection by addressing class imbalance and multimodal complexity. They constructed a balanced dataset called Meme-Merge by merging two existing meme datasets, and implemented a multimodal fusion model combining BERT (baseline and hateXplain variants) with deep ResNet visual encoders. Their model achieved strong performance, with F1-scores of 0.7315 on baseline data and 0.7140 on Meme-Merge, also demonstrating gains in related tasks like metaphor, sentiment, and intention understanding

Recent advances in multilingual meme sentiment analysis [23] have highlighted the importance of building models that generalize across linguistic boundaries. However, most existing works have concentrated on resource-rich languages such as English, with relatively limited exploration of low-resource languages like Bengali, Hindi, or Bodo. Furthermore, the majority of baseline multimodal systems still struggle to consistently outperform strong unimodal baselines [23][24], indicating the need for more effective architectures that can exploit cross-modal dependencies.

3. Dataset

The dataset consists of 378 multimodal Bodo memes, each comprising an image and OCR-extracted text, annotated for five subtasks: sentiment analysis, sarcasm detection, vulgarity detection, abuse detection, and target community identification. Sentiment labels include Positive, Negative, and Neutral, although the present dataset contains only Positive and Negative instances. Sarcasm is categorized as Sarcastic or Non-Sarcastic, while vulgarity is annotated as Vulgar or Non Vulgar. Abuse is labeled as either Abusive or Non-abusive. Target community identification covers categories such as Gender, Religion, Individual, Political, National Origin, Social Sub-groups, Others, and None. The detailed distribution of labels across all five subtasks is presented in Table 1. From the dataset statistics, it is evident that the data distribution is highly imbalanced across multiple subtasks. In the sentiment task, the Neutral class is completely absent, limiting the coverage of sentiment categories. Sarcasm detection is dominated by Sarcastic instances, while abuse detection is skewed toward Non-abusive. Similarly, in target community identification, the majority of samples fall under None and Gender, whereas categories such as Religion, Political, Individual, and Others are extremely underrepresented, and National Origin is missing altogether. Such imbalance poses challenges for training robust models and may lead to biased predictions toward majority classes.

Table 1
Distribution of labels across subtasks in the dataset

Subtask	Labels	Count
Sentiment	Positive	227
	Negative	151
	Neutral	0
Sarcasm	Sarcastic	339
	Non-Sarcastic	39
Vulgarity	Vulgar	107
	Non Vulgar	271
Abuse	Abusive	77
	Non-abusive	301
Target	None	245
	Gender	121
	Social Sub-groups	6
	Individual	3
	Religion	1
	Political	1
	Others	1
	National Origin	0

4. System Architecture

The proposed framework consists of four stages: preprocessing, imbalance handling, training, and testing. In preprocessing, OCR text was tokenized using BERT [25] and images were processed through ResNet [25], with categorical labels encoded for multi-task learning [26]. To address skewed label distribution, RandomOverSampler balanced minority classes. The training stage fused textual and visual features via early fusion, followed by shared layers and task-specific heads for sentiment, sarcasm, vulgarity, abuse, and target prediction. Finally, the trained model was tested on unseen memes, and achieved the predictions. Figure 1 presents the system architecture of the proposed multimodal framework, highlighting the stages of preprocessing, imbalance handling, training, and testing.

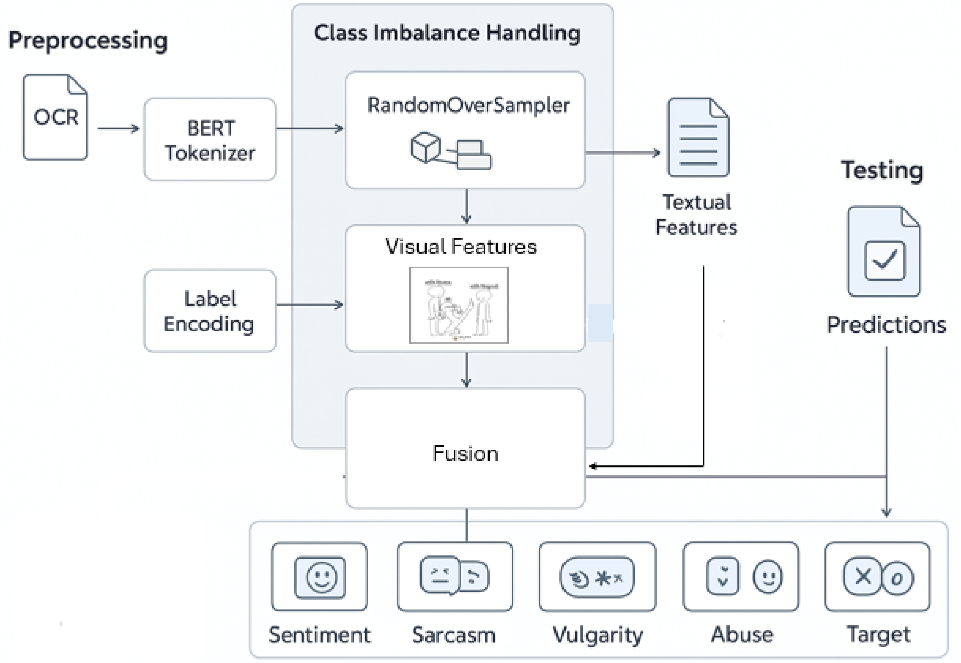

Figure 1: Workflow of the System Architecture

4.1. Preprocessing

Before training the model, several preprocessing steps were applied to ensure that both text and label information were properly structured for multimodal classification. Initially, the missing or irrelevant entries were handled. The textual component of the memes was preprocessed using a BERT tokenizer [25], which converted sentences into subword tokens and subsequently mapped them to numerical IDs suitable for model input. This step ensured consistency in handling vocabulary, including out-of-vocabulary words. For the categorical labels (sentiment, sarcasm, vulgarity, abuse, and target community), label encoding [27] was performed to transform string-based classes into numerical representations. This encoding facilitated multi-label classification [28], where each instance could belong to one or multiple classes simultaneously. Furthermore, the dataset was divided into training and validation splits to enable reliable evaluation of model performance. Through these preprocessing operations, the dataset was standardized into numerical formats for both textual features and categorical labels, making it compatible with the deep learning architecture.

4.2. Class Imbalance Handling

The dataset used in this work exhibits a significant imbalance across multiple classification objectives. Certain categories, for example Positive sentiment or Non-abusive content, occur far more frequently than minority categories such as Sarcastic, Vulgar, Abusive, and Political instances. This imbalance can bias the model toward majority classes and reduce its ability to generalize to rare but crucial cases. To address this issue, we employed the RandomOverSampler technique from the imblearn.over_sampling library [29]. Instead of balancing each task independently, we created a combined label column by concatenating all task-specific labels (Sentiment, Sarcasm, Vulgarity, Abuse, and Target). This ensured that inter-task dependencies were preserved and oversampling operated at the level of unique label combinations.

Formally, if the dataset contained ‘n’ unique combinations of labels, RandomOverSampler adjusted the distribution so that each combination had the same number of samples as the majority combination. For example, if Non-sarcastic, Neutral, Not Vulgar, Non-abusive, None occurred 2,000 times while Sarcastic–Negative, Vulgar, Abusive, Political appeared only 120 times. So, the minority classes ware oversampled through random replication until it reached 2,000 samples. This strategy resulted in a balanced dataset that allowed the multimodal model to learn effectively from both frequent and infrequent categories without being biased toward the majority. Although oversampling increases the dataset size and may introduce redundancy, it was essential for ensuring that the trained models perform well across all subtasks, particularly those involving rare memes.

4.3. Model Training

The multimodal classification framework was trained using a supervised learning approach, integrating both textual and visual modalities in a unified architecture. For the text branch, a BERT-based encoder was employed to process OCR-extracted meme text. Each input sentence was tokenized and transformed into contextual embeddings, which were further fine-tuned during training. The final hidden state of the `[CLS]` token was taken as the condensed textual representation, capturing both semantic and contextual nuances from the meme text. For the image branch, a pre-trained ResNet model [25] was used as the backbone to extract visual features from meme images. The network's convolutional layers captured spatial hierarchies and semantic cues, while the final global average pooling layer produced a compact representation of the visual content. To ensure adaptability to the task, the extracted features were fine-tuned alongside the textual encoder.

After extracting features from both branches, an early fusion strategy was applied. The text and image feature vectors were concatenated to form a joint multimodal representation. This fused vector was then passed through fully connected dense layers with dropout regularization to prevent overfitting. The shared layers facilitated the learning of cross-modal dependencies and improved generalization across diverse meme contexts. The final fused representation was directed to five parallel classification heads, each corresponding to a specific task: sentiment detection, sarcasm detection, vulgarity detection, abuse detection, and target community identification. Each head consisted of a task-specific dense layer with sigmoid activation, enabling the model to produce multi-label predictions simultaneously. The training process followed the following setup:

- **Optimizer:** Adam optimizer with an initial learning rate of 1×10^{-4}.
- **Loss Function:** Binary cross-entropy loss applied independently to each classification head, enabling effective multi-task optimization.
- **Batch Size:** 16 samples per batch.
- **Epochs:** Training was conducted for 5 epochs, with validation performed after each epoch to monitor task-wise performance.
- **Regularization:** Dropout layers were included in the fusion and dense layers to mitigate overfitting.
- **Metrics:** Training and validation accuracy tracked for each classification task.

During training, both text and image inputs were forward-propagated through their respective encoders, fused, and then used to generate task-specific outputs. Gradients from all five classification heads were backpropagated jointly, allowing the shared fusion layers to learn generalized multimodal features. This multi-task learning paradigm not only reduced computational redundancy but also facilitated cross-task knowledge transfer, leading to improved predictions for underrepresented categories. Early stopping was employed based on validation loss to avoid overfitting, and the final optimized model was saved for downstream evaluation on unseen meme samples.

4.4. Testing

The trained multimodal model was applied to the unseen test dataset. For each input pair of meme text and image, the model produced predictions across the five subtasks: Sentiment, Sarcasm, Vulgarity, Abuse, and Target Community. The outputs were mapped to their respective class labels and stored in a structured file, containing the fields: `Ids`, `Sentiment`, `Sarcasm`, `Vulgar`, and `Abuse`. This ensured submission-ready results in the format specified by the organizers.

5. Experimental Results

Table 2
Leaderboard Results for HASOC2025-meme (Bodo)

Rank	Team Name	Score
1	NLPFusion	0.63128
2	FiRC-NLP	0.62217
3	**CNLP-UPES**	**0.60921**
4	SCaLAR	0.60393
5	CSIS BITS Pilani	0.59969
6	CSE_SVNIT	0.58730
7	IIT Dhanbad	0.58186
8	HASOC_2025	0.57776
9	KK_NLP_AI_IIIT_Ranchi	0.57184
10	Golden Ratio	0.56202
11	DeepSemantics	0.56040
12	MUCS	0.55215
13	VEL	0.54566
14	IReL	0.50111
15	HASOC2025-meme (Baseline)	0.39221

5.1. Experimental Setup

The experimental setup utilized a Dell Latitude 5420 laptop running Microsoft Windows 11 Pro (Build 26100). The system was powered by an 11th Gen Intel® CoreTM i5-1135G7 processor (2.40 GHz, 4 cores, 8 threads) with 16 GB RAM and 22.7 GB virtual memory. All preprocessing, training, and evaluation were performed locally on this CPU-based environment without GPU acceleration. The model was trained for 5 epochs using Adam optimizer with a learning rate of 1e-4 and categorical cross-entropy loss. To mitigate class imbalance, RandomOverSampler was employed during training. Performance was evaluated using the official leaderboard metric (macro-averaged F1-score) on the held-out test set.

5.2. Leaderboard Results and Comparative Analysis

Table 2 presents the official leaderboard results for the Bodo track of HASOC2025-meme. Our team, CNLP-UPES, achieved third rank with an F1-score of 0.60921, outperforming several competitive teams, including SCaLAR (0.60393), CSIS BITS Pilani (0.59969), and IIT Dhanbad (0.58186). Although our score was slightly below the top two teams: NLPFusion (0.63128) and FiRC-NLP (0.62217). The margin of difference was relatively small, suggesting that our proposed multimodal fusion strategy is competitive with state-of-the-art approaches. Moreover, our system significantly outperformed the official baseline (0.39221) by a large margin of 21.7% absolute improvement, highlighting the strength of our fusion-based architecture. The comparative analysis reveals three key insights:

- Fusion effectiveness: Experimental results indicate that multimodal fusion approaches consistently outperformed unimodal baselines, with an average relative improvement of approximately 45–55% over the baseline F1 score (0.39221). The superior performance of fusion-based systems validates the hypothesis that the joint representation of OCR-extracted text (semantic cues) and visual embeddings (contextual and stylistic cues) yields a richer feature space, leading to higher discriminative capability in meme classification tasks.
- Model generalization: Despite the inherent limitations posed by the small size and class imbalance of the Bodo meme dataset, our proposed multimodal framework achieved an F1 score of 0.60921, ranking third among fifteen teams. This demonstrates the model's strong generalization ability in low-resource multimodal scenarios, where data scarcity often leads to overfitting. The stability

of results across subtasks further indicates that the architecture effectively learns transferable multimodal representations.
- Performance trade-offs: While the top-ranked team (NLPFusion) achieved a slightly higher F1 score (0.63128, a margin of only 2.2% compared to our model), the relatively narrow performance gap suggests that marginal gains may largely depend on pretraining strategies (e.g., domain-specific language models), fine-grained hyperparameter optimization, or incorporation of auxiliary resources such as external embeddings or larger-scale vision models. This highlights that beyond fusion, optimization at the architectural and training levels can yield incremental improvements.

5.3. Ablation Study

To better understand the contributions of individual components of our framework, we conducted an ablation study [30] using the validation split of the training dataset. The results are summarized below:

- Text-only (BERT embeddings): Achieved an F1-score of 0.542, indicating that textual features alone capture some discriminative signals but are insufficient for complex meme semantics.
- Image-only (ResNet features): Scored 0.498, suggesting that visual information alone provides limited context, especially for memes where meaning is heavily text-dependent.
- Fusion without imbalance handling: When early-fused features were trained without RandomOverSampler, the score dropped to 0.573, demonstrating the critical role of balancing skewed labels.
- Proposed full model (Fusion + Oversampling): Achieved the best performance with 0.609, validating the effectiveness of integrating both modalities and handling imbalance.

6. Conclusion and Future Scope

This work presents a multimodal classification framework for Bodo memes, leveraging OCR-based textual features and visual embeddings to address five key tasks: sentiment detection, sarcasm recognition, vulgarity identification, abuse detection, and target community classification. The experimental findings underscore the critical role of multimodal fusion, where integrating text and image modalities significantly outperformed unimodal baselines. Despite the challenges of data scarcity and imbalance, the system achieved strong generalization capabilities, securing a top-three ranking in HASOC-Meme 2025. The comparative analysis further indicates that while state-of-the-art performance improvements were marginally higher for some systems, the difference can be attributed to optimization factors such as pretraining, fine-tuning, and the use of auxiliary external datasets. This validates the robustness of our proposed model and highlights the potential of multimodal architectures in low-resource contexts.

Looking forward, several avenues can advance this line of research. First, incorporating transformer-based vision-language models (e.g., CLIP, ViLT) may yield stronger cross-modal interactions. Second, extending the dataset through data augmentation and semi-supervised learning can mitigate the limitations of resource scarcity. Third, integrating adversarial training or domain adaptation techniques could enhance robustness across diverse meme genres and platforms. Finally, the expansion of multimodal meme classification to other low-resource languages in Northeast India would not only improve cyber safety but also contribute to inclusive AI systems that safeguard digital spaces for marginalized linguistic communities.

Declaration on Generative AI

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

References

[1] T. Aichner, M. Grünfelder, O. Maurer, D. Jegeni, Twenty-five years of social media: a review of social media applications and definitions from 1994 to 2019, Cyberpsychology, behavior, and social networking 24 (2021) 215–222.

[2] F. Alkomah, X. Ma, A literature review of textual hate speech detection methods and datasets, Information 13 (2022) 273.

[3] J. Habermas, The Theory of Communicative Action, volume 1, Beacon Press, 1984.

[4] G. L. De la Peña Sarracén, P. Rosso, Systematic keyword and bias analyses in hate speech detection, Information Processing & Management 60 (2023) 103433.

[5] H. Zia, et al., Multimodal hate speech detection: Challenges and future directions, ACM Transactions on the Web (2023).

[6] D. Kiela, H. Firooz, A. Mohan, et al., The hateful memes challenge: Detecting hate speech in multimodal memes, in: Advances in Neural Information Processing Systems, 2020.

[7] J. Mao, H. Shi, X. Li, Research on multimodal hate speech detection based on self-attention mechanism feature fusion, The Journal of Supercomputing 81 (2025) 28.

[8] K. Ghosh, M. Das, M. Narzary, S. Saha, S. Barman, A. Mukherjee, S. Modha, D. Ganguly, U. Garain, S. Jaki, T. Mandl, Overview of the hasoc track at fire 2025: Abusive meme identification — shadows behind the laughter, in: K. Ghosh, T. Mandl, S. Pal, S. Majumdar, A. Chakraborty (Eds.), Forum for Information Retrieval Evaluation (Working Notes) (FIRE 2025) December 17-20, Varanasi, India, CEUR-WS.org, 2025.

[9] 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.

[10] M. S. Hee, R. Cao, T. Chakraborty, R. K.-W. Lee, Understanding (dark) humour with internet meme analysis, in: Companion Proceedings of the ACM Web Conference 2024, 2024, pp. 1276–1279.

[11] S. R. Suryawanshi, Unimodal, multimodal and transformational approaches for offensive meme classification: challenges, datasets, and models (2024).

[12] V. Keswani, S. Singh, S. Agarwal, A. Modi, Iitk at semeval-2020 task 8: Unimodal and bimodal sentiment analysis of internet memes, arXiv preprint arXiv:2007.10822 (2020).

[13] M. Hazman, S. McKeever, J. Griffith, Unimodal intermediate training for multimodal meme sentiment classification, arXiv preprint arXiv:2308.00528 (2023).

[14] P. Behera, A. Ekbal, et al., Only text? only image? or both? predicting sentiment of internet memes, in: Proceedings of the 17th International Conference on Natural Language Processing (ICON), 2020, pp. 444–452.

[15] S. Sharma, S. Ramaneswaran, M. S. Akhtar, T. Chakraborty, Emotion-aware multimodal fusion for meme emotion detection, IEEE Transactions on Affective Computing 15 (2024) 1800–1811.

[16] W. Gao, X. Zhao, Emotion classification in internet memes utilizing enhanced convnext and tensor fusion, International Journal of Information Technology (2025) 1–12.

[17] L. Zheng, H. Fei, T. Dai, Z. Peng, F. Li, H. Ma, C. Teng, D. Ji, Multi-granular multimodal clue fusion for meme understanding, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, 2025, pp. 26057–26065.

[18] F. Abdullakutty, U. Naseem, Decoding memes: a comprehensive analysis of late and early fusion models for explainable meme analysis, in: Companion Proceedings of the ACM Web Conference 2024, 2024, pp. 1681–1689.

[19] X. Guo, J. Ma, A. Zubiaga, Nuaa-qmul at semeval-2020 task 8: Utilizing bert and densenet for internet meme emotion analysis, arXiv preprint arXiv:2011.02788 (2020).

[20] A. Avvaru, S. Vobilisetty, Bert at semeval-2020 task 8: Using bert to analyse meme emotions, in: Proceedings of the fourteenth workshop on semantic evaluation, 2020, pp. 1094–1099.

[21] D. S. Chauhan, S. Dhanush, A. Ekbal, P. Bhattacharyya, All-in-one: A deep attentive multi-task

learning framework for humour, sarcasm, offensive, motivation, and sentiment on memes, in: Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing, 2020, pp. 281–290.

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

[23] S. S. Almalki, Sentiment analysis and emotion detection using transformer models in multilingual social media data., International Journal of Advanced Computer Science & Applications 16 (2025).

[24] A. Ranjan, I. Papnai, J. Gupta, M. Aggarwal, A. Saroj, Multilingual detection of persuasion techniques in memes, in: 6th International Conference on Deep Learning, Artificial Intelligence and Robotics (ICDLAIR 2024), Atlantis Press, 2025, pp. 251–260.

[25] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, T. Kong, ibot: Image bert pre-training with online tokenizer, arXiv preprint arXiv:2111.07832 (2021).

[26] Y. Y. Tan, C.-O. Chow, J. Kanesan, J. H. Chuah, Y. Lim, Sentiment analysis and sarcasm detection using deep multi-task learning, Wireless personal communications 129 (2023) 2213–2237.

[27] S. Zhao, X. Hong, J. Yang, Y. Zhao, G. Ding, Toward label-efficient emotion and sentiment analysis, Proceedings of the IEEE 111 (2023) 1159–1197.

[28] A. N. Tarekegn, M. Giacobini, K. Michalak, A review of methods for imbalanced multi-label classification, Pattern Recognition 118 (2021) 107965.

[29] Z. A. Sayyed, Study of sampling methods in sentiment analysis of imbalanced data, arXiv preprint arXiv:2106.06673 (2021).

[30] G. Kasneci, E. Kasneci, Enriching tabular data with contextual llm embeddings: A comprehensive ablation study for ensemble classifiers, arXiv preprint arXiv:2411.01645 (2024).