HASOC-Meme: Enhancing Hate Speech Recognition in Bengali, Hindi, Gujarati, and Bodo Memes Using Multimodal Multitask Transformers

Asha Hegde[1], Sharal Coelho[1] and Amrithkala M Shetty[2]

[1]Department of Computer Science, Mangalore University, India
[2]Department of Computer Applications, Nitte Institute of Professional Education, Nitte (Deemed to be University), Karnataka, India

Abstract

The advancement of memes as a medium for communication on social media has introduced challenges in detecting hate speech and offensive content, particularly in multilingual contexts like India, where languages such as Bengali, Hindi, Gujarati, and Bodo are prevalent. The proposed work is addressed by our team, NLPFusion, by focusing on identifying hate speech in memes across these four Indic languages. This study offers a multimodal multitask transformer framework to enhance hate speech recognition in memes, integrating advanced vision and language models. For Bangla memes, we employ ConvNeXt-small, a state-of-the-art convolutional architecture inspired by transformers, to extract robust semantic visual features. ResNet-34 is used for Hindi and Bodo memes, while ResNet-18 is applied for Gujarati memes. Textual data is processed using transformer-based models with domain-specific pre-training to handle linguistic diversity and code-mixed content. Evaluated on the HASOC-Meme dataset, the model achieved macro F1-scores of 0.634 (Gujarati, 2nd), 0.631 (Bodo, 1st), 0.623 (Hindi, 2nd), and 0.608 (Bangla, 5th). These results demonstrate the framework's effectiveness in addressing the complexities of multimodal, multilingual hate speech detection, offering a scalable solution for safer online environments.

Keywords

Hate Speech, HASOC, Multimodal, Multitask Learning, Transformer

1. Introduction

Social media sites like Twitter and Facebook have become hugely popular because they are so easy to use and are so extensively available, providing users with an influential platform for articulating their opinions [1]. Users across all age groups are actively participating on these sites, regularly documenting and sharing aspects of their everyday lives, which adds to an ever-increasing amount of user-generated content. While there are numerous benefits to social media, it is not without its disadvantages. A considerable amount of offensive and damaging content—such as hate speech—is available online, which has serious societal implications [2]. Online toxic content can compromise democratic processes. Most social media platforms have started actively policing the content their users share. This development has necessitated a rising demand for automated systems that can identify and mark as suspicious or potentially harmful posts [3]. Therefore, online societies, technology firms, and social media platforms are making significant investments in software and technologies to detect and govern abusive speech in order to create more secure online spaces.

The detection of hate speech more and more requires the examination of multimodal data, since toxic online content tends to take advantage of the integration of text and images to express hateful messages in veiled or coded ways. In most cases, the textual information by itself can look harmless, without any overt signs of hate or offense. Likewise, a linked image, when examined independently, can look harmless or unclear. When presented separately, however, these two modalities may provide a powerful yet deeply offensive composite message. Examining both text and image content at the same time allows for a better understanding of the environment in which hate speech is embedded [4].

Forum for Information Retrieval Evaluation, December 17-20, 2025, Varanasi, India
hegdekasha@gmail.com (A. Hegde); sharalmucs@gmail.com (S. Coelho); amrithkalas@gmail.com (A. M. Shetty)
© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-02-24

CEUR-WS.org/Vol-4173/T9-7.pdf

Multimodal analysis can pick up on subtle or hidden hateful messages that may otherwise pass undetected. Hence, merely the textual data cannot be sufficient to judge them as images may provide additional context in order to make an accurate judgement. Hence, it is crucial to build systems which are capable of interpreting and processing multimodal inputs properly in order to detect online hate speech accurately [5]. These systems can dramatically improve the reliability and accuracy of content moderation tools to identify and deal with offensive content accurately in a timely manner. This is essential in developing safer and more welcoming virtual spaces. In the process of stifling the dissemination of toxic, injurious, and hateful content on social media and virtual societies, multimodal hate speech detection ultimately helps safeguard democratic dialogue, user wellness, and social concord in the digital world.

In this paper we employed domain-specific pre-training to handle the linguistic diversity and cultural nuances of the target languages, ensuring effective processing of code-mixed text and contextually rich visual content. This work aims to contribute to safer online environments by providing an efficient and scalable solution for automated hate speech detection in multilingual memes, addressing a critical need in the context of India's diverse linguistic landscape.

2. Related Work

Hate speech detection is a recognized research issue among researchers of different languages [6][7]. Most of the hate speech detection systems are developed based on textual data, which is collected from social media and other digital resources. Some works are also performed concerning the low-resource languages [8][9] [9]. Earlier researchers widely used Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) Networks, and the combination of RNN and Convolutional Neural Network (CNN) based methods. In contrast to the text-based analysis, in recent years, few pieces of work have considered multimodal information (i.e., image + text) for hate speech detection. Kiela et al.[10] presented a multimodal memes dataset for detecting hate speech. In another work,

Several approaches are employed for detecting hate speech using multimodal learning. Rana and Jha [11] introduced a multimodal hate speech dataset concerning three modalities (i.e., image, text, and audio). Similarly, Karim et al. [12] created a dataset for multimodal hate speech detection from Bengali memes. Some researchers exploited the different fusion techniques (i.e., early and late fusion) to evaluate the image and textual features jointly.

3. Task Description

The HASOC 2025 shared Task provide meme for abuse, sentiment, sarcasm, vulgarity detection. The task is to analyze multimodal data (text and image) in order to detect abuse, find targeted communities, measure vulgarity and sarcasm, and determine sentiment labels [13], [14]. Therefore, the task will be divided into five parts.

1. Sentiment detection: It is a multi-class classification task containing 3 labels: Positive, Neutral, and Negative. The meme is imparting a positive, funny, or grateful tone / or it is not extremely positive or negative in tone / or the meme conveys hostility, mockery, or criticism.
2. Sarcasm Detection: It is a binary classification task containing 2 classes, namely, Sarcastic and Non-Sarcastic. The meme offers statements or images that convey the opposite of their literal sense, sometimes to ridicule or scoff, and is labelled as Sarcastic. The meme clearly communicates its message in a non-sarcastic or non-ironic manner and is labelled as Non-Sarcastic.
3. Vulgarity Detection: It is a binary classification task containing 2 classes, namely, Vulgar and Not Vulgar. Vulgar - The meme has offensive or insulting words, gestures, or images. Not Vulgar - The meme does not carry any such material.
4. Abuse Detection: It is a binary classification task containing 2 classes, namely, Abusive and Not abusive. Abusive - The meme carries offensive, hurtful, or insulting language, images, or

suggestion against a person or a group of people. Not abusive - The meme has no abusive, offensive, harmful, or pejorative content.

4. Methodology

In the proposed work few computational models have been explored to identify hateful memes by considering the combination of image and text modalities. This section briefly discusses the methods and parameters utilized to construct the models.

4.1. Pre-processing

The methodology includes preprocessing for the two modalities–text and images. Each requires specific pre-processing techniques, as outlined below.

(a) **Image Preprocessing** : In order to improve the visual encoder's generalization capability, the following series of augmentation methods were used on images while training:

- Random Resized Crop: Randomly resizes and crops the image to 224×224 with the scaling factor ranging between 0.8–1.0 to mimic diverse zoom factors.
- Random Horizontal Flip: Adds left-right flipping to enhance orientation invariance.
- Color Jitter: Randomly alters brightness, contrast, and saturation by ±0.2.
- Random Rotation (±15°) and Affine Transformations: Provides strength to geometric distortions.
- Normalization: Images are normalized at last with ImageNet statistics.

(b) **Text Preprocessing** Optical Character Recognition (OCR) text extracted was pre-processed to minimize noise:

- Indic Normalization: Bangla text was normalized with the Indic NLP Library (if available), which manages canonical variations and cleans from inconsistencies in Unicode representations.
- Noise Removal: Special characters and duplicate punctuation were deleted, without losing strong sentiment markers such as ! and ?.
- Lowercasing and Stripping: Text was put into lowercase and stripped of extra spaces for uniformity.

4.2. Text Encoder

The text encoder is responsible for encoding the semantic and contextual information of the input text. The raw text is first tokenized into subword units, enabling the model to represent rare words and cope with morphologically complex language effectively. Each token is later embedded into a high-dimensional space and subjected to several layers of a transformer-based model, which employs self-attention to capture long-range dependencies and contextual relationships between words.

- **Bangla** - For Bangla Language, we used XLM-RoBERTa Large[1], a transformer-based multilingual language model pre-trained on massive CommonCrawl corpora. It employs subword tokenization, allowing efficient processing of Bangla's rich morphology and intricate word constructions. The input sequence is encoded as contextual embeddings, and the special classification token is taken out to capture the global semantic sense. Fine-tuning upper layers guarantees adaptation to Bangla-specific factors like script differences and contextual sentiment indicators.
- **Hindi** - We employed Hindi-RoBERTa, a transformer model pre-trained on large-scale Hindi corpora, for Hindi. Its monolingual nature enables it to acquire detailed linguistic and semantic properties particular to Hindi, such as dealing with compound words, gender markers, and honorifics. With the use of strong contextual embeddings, the encoder yields robust representations for Hindi social media text that may contain code-mixed and colloquial content.

[1]https://huggingface.co/FacebookAI/xlm-roberta-large

- **Gujarati**[2] - For Gujarati, we used Gujarati-BERT, a transformer model that has been specifically pre-trained on Gujarati text. The model is aided by a subword-level tokenizer that effectively deals with Gujarati's script and morphological richness. Its embeddings capture syntactic as well as semantic information and thus perform well in modeling subtle expressions, idiomatic uses, and sentiment-carrying indicators frequently encountered in Gujarati discourse. Fine-tuning further brings the encoder in line with downstream classification tasks.
- **Bodo** - For Bodo, we used Multilingual BERT (mBERT), which is trained on 104 languages' Wikipedia texts. With its shared subword vocabulary, it generalizes to low-resource languages such as Bodo, in which there are no large annotated corpora. The encoder generates contextual embeddings that reflect sentence-level semantics, and fine-tuning enables it to adjust to the particular syntactic and semantic characteristics of Bodo text. This makes it suitable for tasks such as sentiment and abuse detection in resource-limited settings.

4.3. Image encoder

For Bangla memes, we used ConvNeXt-small, a state-of-the-art convolutional architecture motivated by transformer design yet refined for CNN efficacy. ConvNeXt uses depthwise convolutions, deep kernel sizes, and layer normalization to create extensive hierarchical representations of visual information. By adjusting fine-grained upper layers on the meme dataset without keeping pretrained lower-level filters, the encoder derives strong semantic features such as objects, context, and backgrounds that help form the sentiment or intent of the meme.

We employed ResNet-34, a residual network architecture with 34 layers, for Hindi memes. ResNet's skip connections address the vanishing gradient problem, allowing deeper feature learning without loss of performance. The encoder successfully extracts both low-level (edges, textures) and high-level (objects, scenes) visual features, which are essential in inferring humor, sarcasm, or abuse hidden in the images.

For Gujarati memes, we used ResNet-18, a light residual network of 18 layers. Although less deep than ResNet-34, it offers adequate feature extraction strength for the comparatively smaller dataset size. The encoder retains key visual patterns while providing computational frugality, allowing for quicker training and lowered danger of overfitting on resource-limited Gujarati meme data.

For Bodo memes, we also used ResNet-34 as the image encoder. The structure of this architecture has a good balance between computational power and strong representational ability, hence appropriate for medium-resource tasks. Its hierarchical residual blocks allow salient features, like characters and symbolic hints, to be extracted from an image, enhancing textual knowledge in memes for proper classification.

4.4. Fusion strategy

Multimodal fusion of text and image features was conducted using the late fusion strategy. The text encoder and image encoder's output embeddings were first used. To be compatible with various architectures and dimensionalities, both the embeddings were projected into one shared latent space via fully connected layers. After being aligned, the text and image representations projected were concatenated to create a shared multimodal feature vector. The combined representation was then fed through a feed-forward projection layer and onto ReLU and dropout ($p = 0.5$) for overfitting prevention and generalization improvement. This approach enables the model to integrate complementary information from the textual and visual modalities efficiently, supporting strong classification in any language and meme sets.

[2]https://huggingface.co/l3cube-pune/gujarati-bert

4.5. Model Building

4.5.1. Multi-task learning

The fused representation was passed to four parallel classification heads for multi-task learning:

- Sentiment Classification: 3 classes (Positive, Neutral, Negative)
- Sarcasm Detection: 2 classes (Sarcastic, Non-Sarcastic)
- Vulgarity Detection: 2 classes (Vulgar, Non Vulgar)
- Abuse Detection: 2 classes (Abusive, Non-abusive)

Table 1
Hyperparameters used in the multimodal experiments.

Hyperparameter	Value
Tokenizer max length	128
Image input size	224×224
Batch size	16
Number of epochs	20
Optimizer	AdamW
Learning rate	2e-5
Weight decay	0.01
Scheduler	Linear warmup + decay
Warmup steps	10% of total steps
Loss function	Focal Loss ($\alpha = 0.25,\ \gamma = 2.0$)
Gradient clipping	max_norm = 1.0
Gradient accumulation	2 steps
Dropout	0.5
Fusion projection size	512
Random seeds	42, 123, 456
Model selection metric	Average Macro F1 (across 4 tasks)

4.5.2. Loss Function and Optimization

A Focal Loss with class-specific weights was used to handle severe class imbalance. Dynamic task-specific loss weighting was applied during training, where tasks with lower F1 scores received higher weights in subsequent epochs. AdamW with differential learning rates (1e-5 for encoders, 2e-5 for classification layers) is used to build the model. Further, Learning Rate Scheduler (Linear Warmup + Decay) is employed. Gradient Clipping and gradient accumulation were used to stabilize training under limited GPU memory. The additional hyperparameters used in the multimodal experiments are shown in Table1

5. Experiments and Result

The experiments were conducted using the HASOC-Meme dataset, which includes memes in Bengali, Hindi, Gujarati, and Bodo, each comprising text and image modalities. The model was implemented using PyTorch and the Hugging Face Transformers library.

To ensure reproducibility and robustness, experiments were conducted with multiple random seeds (42, 123, and 456). For each configuration, the models were trained and evaluated across all four downstream tasks. The best-performing model was selected based on the average macro F1 score, providing a balanced measure of performance across different classes. This selected model was then used to generate predictions on the final test sets, ensuring consistency and reliability in the reported results.

Table 2
The performance of the proposed models based on macro F1-scores and corresponding ranks

Ranks Obtaned	Language	Result
1	Bodo	0.63128
2	Hindi	0.62398
2	Gujarati	0.63436
5	Bangla	0.60834

The performance of the proposed multimodal multitask transformer model on the HASOC-Meme dataset is summarized in the table 2, based on macro F1-scores and corresponding ranks in the shared task.

The model achieved the highest performance on Gujarati with F1-score of 0.63436 and Bodo obtained F1-score of 0.63128, securing 2^{nd} and 1^{st} ranks, respectively, in the HASOC-meme shared task. Hindi memes yielded a strong F1-score of 0.62398, also ranking 2^{nd}. Bangla memes, despite the advanced ConvNeXt-small architecture, obtained an F1-score of 0.60834, placing 5^{th}, likely due to the complexity of semantic features in Bangla meme imagery. The results highlight the effectiveness of tailored image encoders (ConvNeXt-small, ResNet-34, ResNet-18) and the multitask framework in handling the linguistic and visual diversity of multilingual memes.

6. Conclusion

This study introduced a multimodal multitask transformer framework for hate speech and offensive content identification in the HASOC-meme dataset, covering Bengali, Hindi, Gujarati, and Bodo memes. By incorporating advanced image encoders such as ConvNeXt-small for Bangla, ResNet-34 for Hindi and Bodo, and ResNet-18 for Gujarati along with a fine-tuned IndicBERT text encoder, the model effectively grasped the interplay of visual and textual cues in multilingual memes. The multitask learning approach, optimizing for hate speech detection, offensive content classification, and sentiment analysis, improved generalization across various linguistic and cultural contexts. The model achieved top ranks in the HASOC-meme shared task, with F1-scores of 0.63436 (Gujarati, 2nd), 0.63128 (Bodo, 1st), 0.62398 (Hindi, 2nd), and 0.60834 (Bangla, 5th). These results highlight the framework's ability to address the complexities of multimodal hate speech detection in low-resource Indic languages.

Declaration on Generative AI

In preparing this work, the author(s) utilized Grok[3] for grammar and spelling checks. Paraphrasing was handled via QuillBot. With this tool, the author(s) reviewed and revised the content as required, while assuming full responsibility for the publication's integrity.

References

[1] S. Coelho, A. Hegde, H. L. Shashirekha, et al., Mucs@ lt-edi2023: Detecting signs of depression in social media text, in: Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion, 2023, pp. 295–299.

[2] W. Yin, A. Zubiaga, Towards generalisable hate speech detection: a review on obstacles and solutions, PeerJ computer science 7 (2021) e598.

[3] A. Hegde, H. Shashirekha, Transformer-driven multi-task learning for fake and hateful content detection, in: Proceedings of the 21st International Conference on Natural Language Processing (ICON): Shared Task on Decoding Fake Narratives in Spreading Hateful Stories (Faux-Hate), 2024, pp. 29–35.

[3] https://grok.com

[4] N. Marutyan, A. Jouljian, The hateful memes challenge: Detecting hate speech in multi-modal memes (2024).
[5] A. Naseeb, M. Zain, N. Hussain, A. Qasim, F. Ahmad, G. Sidorov, A. Gelbukh, Machine learning-and deep learning-based multi-model system for hate speech detection on facebook, Algorithms 18 (2025) 331.
[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: Proceedings of the 15th annual meeting of the forum for information retrieval evaluation, 2023, pp. 13–15.
[7] F. Alkomah, X. Ma, A literature review of textual hate speech detection methods and datasets, Information 13 (2022) 273.
[8] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, A. Mukherjee, Hatexplain: A benchmark dataset for explainable hate speech detection, in: Proceedings of the AAAI conference on artificial intelligence, volume 35, 2021, pp. 14867–14875.
[9] M. Das, A. Mukherjee, Banglaabusememe: A dataset for bengali abusive meme classification, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 15498–15512.
[10] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, Advances in neural information processing systems 33 (2020) 2611–2624.
[11] A. Rana, S. Jha, Emotion based hate speech detection using multimodal learning, arXiv preprint arXiv:2202.06218 (2022).
[12] M. R. Karim, S. K. Dey, T. Islam, M. Shajalal, B. R. Chakravarthi, Multimodal hate speech detection from bengali memes and texts, in: International Conference on Speech and Language Technologies for Low-resource Languages, Springer, 2022, pp. 293–308.
[13] 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, 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.
[14] 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.