Overview of the HASOC Track at FIRE 2025: Abusive Meme Identification — Shadows Behind the Laughter

Koyel Ghosh[1,2], Mithun Das[3], Shubhankar Barman[4], Mwnthai Narzary[5], Saptarshi Saha[6], Animesh Mukherjee[3], Sandip Modha[7,8], Debasis Ganguly[9], Utpal Garain[6], Sylvia Jaki[1,10] and Thomas Mandl[1,7]

[1]University of Hildesheim, Hildesheim, Germany
[2]SRM Institute of Science & Technology, Kattankulathur, India
[3]Indian Institute of Technology, Kharagpur, India
[4]BITS Pilani, India
[5]Central Institute of Technology, Kokrajhar, India
[6]Indian Statistical Institute, Kolkata, India
[7]Dhirubhai Ambani University, Gandhinagar, India
[8]University of Milano-Bicocca, Milan, Italy
[9]University of Glasgow, United Kingdom
[10]Katholieke Universiteit Leuven, Campus Antwerp, Belgium

Abstract

The dramatic surge in the use of social media platforms for sharing information and opinions has, unfortunately, also fueled a sharp rise in online abuse. One of the simplest yet most insidious tools for such abuse is the meme: a visual artifact that typically fuses an image with a short, often provocative, text overlay. While memes can be humorous or satirical, they are increasingly being weaponized to target individuals or communities through hateful or derogatory content. This poses a serious threat to online safety and digital well-being. Detecting and curbing such abusive memes is therefore an urgent necessity. However, the task becomes significantly more challenging in low-resource Indian languages such as Bangla, Hindi, Gujarati, Bodo etc., where annotated benchmark datasets are scarce or entirely absent. Without such datasets, the development and evaluation of robust AI models remain severely constrained. In this work, we aim to bridge this crucial gap by constructing a multilingual abusive meme dataset—covering Bangla, Hindi, Gujarati, and Bodo, thereby enabling research and innovation in abusive meme detection across low-resource Indian languages. Each dataset is annotated with five classification labels: sentiment, sarcasm, vulgarity, abuse, and target. A total of 20 unique teams participated, submitting over 306 runs across the four languages. The performance of the best systems was evaluated using the Macro F1 score, with the top-performing models achieving scores of 0.6275 (Bangla), 0.6570 (Hindi), 0.6750 (Gujarati), and 0.6312 (Bodo). This article briefly summarizes the tasks, data development, results and approaches.

Disclaimer: This paper includes certain content that may be considered offensive; however, its inclusion is necessary due to the nature of the work.

Keywords

Hate Speech, Abusive Meme Identification, Social NLP, Social Media, Memes, Deep Learning, Low-Resource Language, Indian Language, Benchmark, Bangla, Hindi, Gujarati, Bodo, HASOC

1. Introduction

The concept of memes was first proposed by Richard Dawkins in 1976, describing them as cultural elements that spread from person to person much like genes, quickly sharing ideas and influencing group thinking [1]. Yet, memes can also pose certain risks that deeply affect both people and society

Forum for Information Retrieval Evaluation, December 17-20, 2025, India
koyelg@srmist.edu.in (K. Ghosh); mithun.rcciit@gmail.com (M. Das); contact.shubhankarbarman@gmail.com (S. Barman); ph22cse1004@cit.ac.in (M. Narzary); Saptarshi2016saha@gmail.com (S. Saha); animeshm@gmail.com (A. Mukherjee); sjmodha@gmail.com (S. Modha); debforit@gmail.com (D. Ganguly); utpal.garain@gmail.com (U. Garain); sylvia.jaki@kuleuven.be (S. Jaki); mandl@uni-hildesheim.de (T. Mandl)
0000-0001-5347-4961 (K. Ghosh); 0000-0003-1442-312X (M. Das); 0000-0003-4534-0044 (A. Mukherjee); 0000-0003-2427-2433 (S. Modha); 0000-0003-0050-7138 (D. Ganguly); 0000-0001-7207-5018 (U. Garain); 0000-0001-7840-7300 (S. Jaki); 0000-0002-8398-9699 (T. Mandl)

as a whole. In today's world, "Internet memes" or "image memes" are widely seen on social media platforms. A meme is typically a visual idea that combines an image with a short caption written over it, making the text an essential part of the picture [2]. Internet memes have become powerful means of communication and self-expression, quickly spreading ideas, feelings, and cultural meanings across online communities. While they are often created for fun and humor, memes also play an important role in shaping public opinion and political discussions. Although usually intended as jokes, in today's online world, memes go far beyond simple humor. People create them easily and at no cost, using them to attract attention and gain social recognition. The growing presence of memes on social media has raised both curiosity and concern about their influence on society, as they play a major part in digital culture and mirror the shared mindset of online communities. Malicious users frequently employ memes to intimidate or harm individuals or specific groups. This form of content is termed abusive memes. Owing to their rapid shareability, such memes can escalate social conflicts, harm the credibility of online platforms [3], and cause significant emotional distress to affected individuals [4]. Hence, it is essential to regulate their dissemination, and the primary step in this direction is the accurate identification of abusive memes. Memes have been found to be highly effective in spreading hateful ideologies, due to their seemingly humorous undertone. Research has shown that hateful memes can be surprisingly persuasive, often going viral and resonating with a wide audience [5]. This phenomenon highlights the need to examine the role of memes in perpetuating hateful ideologies and their potential impact on society. Memes can develop a highly persuasive effect due to their seemingly humorous undertone [6]. In recent years, several studies have attempted to detect and mitigate the impact of abusive memes on social media. However, most of this research has focused primarily on memes containing English text [7, 8, 9, 10, 11]. Moreover, multiple multimodal vision language models have been investigated, but most of them remain confined to English. Research efforts in other languages are still limited, especially for low-resource Indian languages such as Bangla, Hindi, Gujarati, Bodo, etc.

The HASOC (Hate Speech and Offensive Content Identification) track, which has provided a platform for hate speech detection since 2019 at FIRE (Forum for Information Retrieval Evaluation) [12]. This year, HASOC 2025 provides a comprehensive platform for advancing research on hate speech and offensive content across multiple modalities and languages. In this context, the present paper offers an overview of abusive meme identification datasets available in Bangla, Hindi, Gujarati, and Bodo. These datasets are crucial, as they contribute valuable task-specific resources for languages that are traditionally low-resourced. By compiling and analyzing these meme datasets, the paper aims to support the development of more robust and inclusive multimodal hate speech detection systems. Here, we present the task, dataset, annotation process, participants' results, as well as their short system descriptions.

2. Related Work

2.1. Multimodal Abusive Meme Datasets

Research on multimodal abusive or hateful memes has grown steadily, supported by the development of several datasets over the last few years. One of the earliest initiatives was by Sabat et al. [13], who collected 5,020 memes from Google Images to study hate meme detection. This was followed by the large scale MMHS150K dataset by Gomez et al. [14], consisting of 150K meme like posts gathered from Twitter. Chandra et al. [15] broadened the scope by constructing a dataset dedicated to antisemitic memes using Twitter and Gab. Likewise, Suryawanshi et al. [16] compiled 743 election related memes annotated for offensiveness, while Pramanick et al. [17] introduced a dataset of around 3.5K harmful COVID-19 memes.

A major milestone in this space is the Hateful Memes Challenge dataset created by Facebook AI [18]. It marked the beginning of large-scale, carefully curated multimodal hate speech datasets. The authors emphasized the context-dependent nature of memes: an identical image paired with different text may or may not be hateful. To capture this phenomenon, the dataset includes manually constructed confounders, in which either the image or the text was replaced with a benign counterpart. This

prevents models from exploiting superficial correlations, although it also results in memes that may not naturally occur on social media.

The release of this dataset catalyzed a series of multimodal shared tasks. SemEval-2022 Task 5 (MAMI) [19] focused on misogynistic multimodal content, while EXIST 2024 [20][1] explored abusive and aggressive communication across languages. Pandiani et al. [21] offer an overview of 34 multimodal toxic content datasets, highlighting that most resources remain English centric and that many world languages lack sufficient multimodal hate-speech data.

2.2. Datasets for Indian Languages

In contrast to English, multimodal meme datasets for Indian languages are limited and often not publicly available. Karim et al. [22] expanded the Bengali hate speech dataset [22] by labelling 4,500 memes, but the dataset is not publicly released, and details about data collection, annotation protocols, target communities, and inter-annotator agreement are lacking. Hossain et al. [23] developed another Bengali dataset containing 4,158 memes with Bengali and code-mixed captions. However, this dataset too remains unavailable, and critical labels, such as target groups are missing. Das et al. [24], their dataset contains 4,043 Bengali memes. Among them, 1,515 are tagged as abusive and 2,528 as non-abusive. A total of 1,664 memes are marked as sarcastic, while 2,379 are not. For vulgarity, 1,171 memes are labeled vulgar and 2,872 non-vulgar. In terms of sentiment, 592 express positive sentiment, 1,414 are neutral, and 2,037 convey negative sentiment. These limitations echo the concerns raised by Kirk et al. [25], who note that memes in research datasets often fail to represent the diversity, stylistic variation, and informality found in real social media memes.

Emoffmeme [26] is a Hindi meme dataset comprising 7,500 samples designed for offensive meme detection, where offensiveness is annotated using a binary (yes/no) labeling scheme. In addition, the dataset captures the emotional or sentiment aspect of memes through a multi-label, multi-class framework, covering emotions such as fear, neglect, irritation, rage, disgust, nervousness, shame, disappointment, envy, suffering, sadness, joy, pride, and surprise. The Indian Political Memes (IPM) dataset [27] contains 1,218 Hindi and English memes focused on hateful meme detection. Unlike binary schemes, IPM adopts a single-label multi-class annotation distinguishing between non-offensive, hate-inducing, and satirical memes. MultiBully [28] includes 5,854 Hindi and English memes for cyberbullying detection, while MultiBully-Ex [29] extends it with 3,222 memes focused on cyberbullying explanation. Both datasets annotate bullying using binary labels (bully/non-bully) and further apply single-label multi-class schemes for emotions (e.g., joy, sadness, fear, anger, disgust, surprise, anticipation, trust, and ridicule) as well as coarse sentiment categories (positive, neutral, negative). In addition, they classify the degree of harmfulness into very harmful, partially harmful, and harmless using a single-label multi-class setup, enabling a finer-grained analysis of harmful content. Irony and sarcasm are annotated in binary form in the MultiBully [28] and MultiBully-Ex [29], respectively. Pol_Off_Meme [30] is a Hindi and English dataset of 7,500 memes for offensive meme detection. It uses binary labels (yes/no) to mark offensiveness, further distinguishes implicit and explicit offensiveness through binary annotation, and captures emotions via a multi-label multi-class scheme covering fear, neglect, irritation, rage, disgust, nervousness, shame, disappointment, envy, suffering, sadness, joy, pride, and surprise; political attributes are also identified using a binary label indicating whether a meme is political or not.

TamilMemes [31] is a Tamil-language dataset containing 2,969 memes created for troll meme detection. In this dataset, trolling is annotated as a dedicated dimension using a binary labeling scheme (yes/no).

While most existing efforts focus on English, abusive meme detection in Indian languages remains largely underexplored, with the exception of Bengali and Hindi. To bridge this gap, our work presents a multimodal abusive-meme dataset for Bangla, Hindi, Gujarati, and Bodo, designed to closely reflect memes typically shared on real social media platforms. The scarcity of datasets, limited transparency, and lack of resources that capture real-world meme distributions strongly motivate the development of high-quality datasets for Indian languages, which forms the core motivation of this study.

[1] https://nlp.uned.es/exist2024/ (Access on 05.12.2025)

3. Task Description

In HASOC meme 2025, the task is with four languages proposed in the research area of hate speech detection. These tasks offered all four languages: Bangla, Hindi, Gujarati, and Bodo. Figure 1 shows the Screenshot of HASOC meme Website [2].

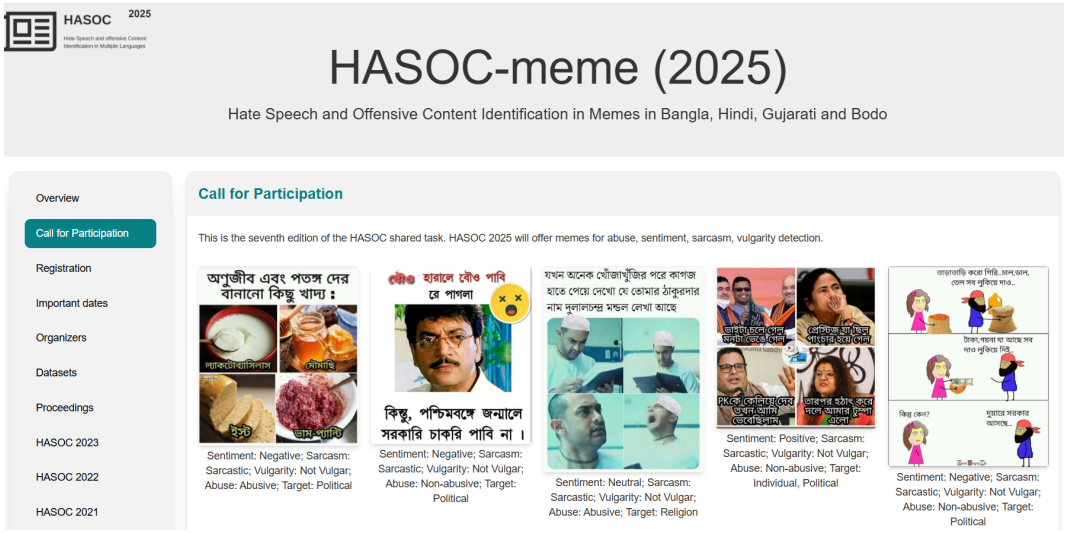

Figure 1: Screenshot of HASOC-meme 2025 Website

This task involves analyzing multimodal data (image and text) to detect abuse, identify targeted communities, assess vulgarity and sarcasm, and assign sentiment labels. So, the task will be in five parts. (1) Sentiment detection: positive, negative, and neutral, (2) Sarcasm Detection: sarcastic or not, (3) Vulgarity Detection: vulgar or not, (4) Abuse Detection: abusive or not, and (5) Target Community Identification: Gender, Religion, Individual, Political, National Origin, Social Sub-groups, Others, and none.

4. Dataset Description

In this section, dataset collection, annotation, and analysis have been discussed for HASOC-meme.

4.1. Dataset Collection

Our primary aim in constructing this multilingual, multimodal dataset is to ensure its diversity, so we intentionally selected some Gender, Religion, Political, communal ™Facebook meme pages, ™Instagram, ™Google Image, ™Bing, and ™YouTube channels etc. Unlike the Hateful Memes Challenge [18], which used synthetically generated memes, our dataset is built entirely from real world memes. After collecting the images, we performed several filtering steps before annotation. Memes without text, memes containing text in languages other than the targeted four languages, and memes with very low resolution that made the text unreadable were removed. These same preprocessing steps were consistently applied across all four languages: Bangla, Hindi, Gujarati, and Bodo to ensure clean and reliable datasets.

4.2. Dataset Annotation

For the annotation process, we engaged fourteen undergraduate students (aged 20–25 years), consisting of nine males and five females, to identify whether each meme was abusive or non-abusive. These

[2] https://hasocfire.github.io/hasoc/2025/call_for_participation.html (Access on 05.12.2025)

Annotation Type	Objective	Labels / Description
Sentiment	Assign a sentiment label to the meme	**Positive:** Supportive, humorous, or appreciative tone **Neutral:** Neither positive nor negative **Negative:** Hostility, mockery, or criticism
Sarcasm	Determine whether the meme conveys sarcasm or irony	**Sarcastic:** The meme conveys sarcasm or irony **Non-Sarcastic:** The meme does not convey sarcasm or irony
Vulgarity	Identify vulgar or obscene language/imagery	**Vulgar:** Explicit or offensive words/gestures **Not Vulgar:** No explicit or offensive content
Abuse	Determine whether the meme is abusive	**Abusive:** Harmful, offensive, or derogatory content targeting individuals or groups **Non-Abusive:** No harmful or derogatory content
Target Community	Identify the community targeted by the meme, if any	**Gender:** Male, female, non-binary, transgender **Religion:** Religious beliefs, deities, practices **Individual:** Targeting a specific person. **Political:** Political ideologies, parties, people **National Origin:** Country or ethnicity-based groups **Social Sub-groups:** Socio-economic, cultural, occupational groups **Others:** Any target not in above categories **None:** No specific target

Table 1
Annotation guidelines for the six annotation types, along with their objectives and descriptions, for Bangla, Hindi, Gujarati, and Bodo languages.

annotators worked across all four languages, like Bangla, Hindi, Gujarati, and Bodo, and were supported with OCR-extracted text. All students were native speakers of the respective languages and were compensated fairly according to local standards. The entire annotation workflow was overseen by a post-doctoral researcher, Ph.D. researchers, and an industry-academic researcher, each with several years of experience in analyzing harmful social media content. We adopt the same schemes as the authors [24] for the annotation rules with little necessary updates. Our framework includes six types of annotations, as presented in the table 1 below.

4.3. Dataset Analysis

Figure 2 shows the train and test split of the Bangla, Hindi, Gujarati, and Bodo datasets. We summarize the key statistics of our multilingual meme traning datasets in Table 2. Across all four languages, noticeable label imbalance is present in every annotation category. In the Bangla dataset, negative sentiment (1,476 samples) is substantially higher than positive (906) and neutral (311) classes. Hindi exhibits a similar pattern, where negative sentiment dominates the other two classes. In the Gujarati, more positive and neutral labels are present than negative. The Bodo dataset is particularly imbalanced, containing no neutral samples and showing a clear skew toward the positive class (227 positive vs. 151 negative).

A strong imbalance is also observed in sarcasm labels. Bangla contains a significantly higher proportion of sarcastic memes (2,081) compared to non-sarcastic ones (612). Hindi (770 vs. 371) and Gujarati (670 vs. 219) show similar skewed distributions. In the Bodo dataset, this imbalance becomes even more pronounced, with 339 sarcastic and only 39 non-sarcastic samples.

Vulgarity labels show imbalance in the opposite direction for most languages. All four datasets have considerably more non-vulgar than vulgar samples. For instance, Bangla (2,226 non-vulgar vs. 467 vulgar), Hindi (764 vs. 377), Gujarati (592 vs. 297), and Bodo (271 vs. 107), indicating that vulgar memes represent a minority class.

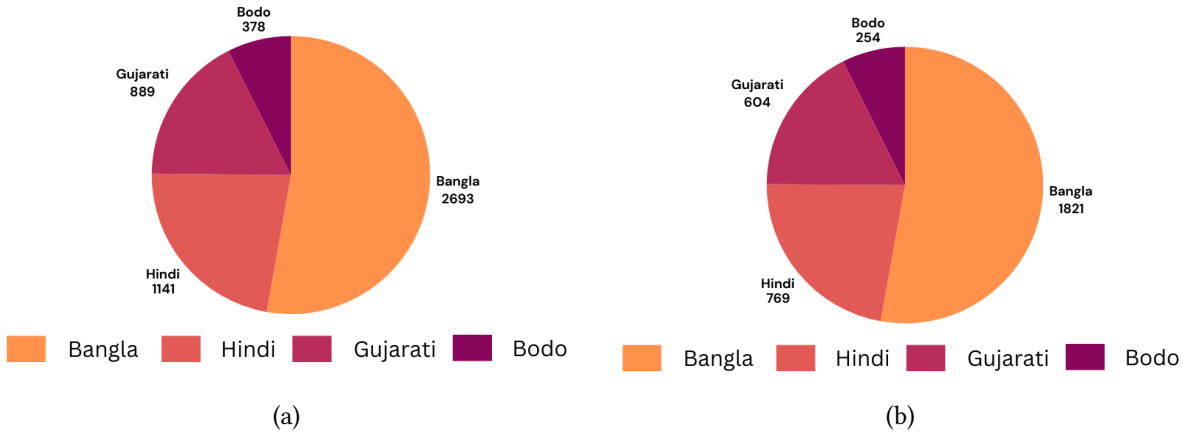

Figure 2: Train (a) - Test (b) split of the Bangla, Hindi, Gujarati, and Bodo datasets.

Language	Sentiment			Sarcasm		Vulgarity		Abuse	
	Positive	Negative	Neutral	Non-Sarcastic	Sarcastic	Non Vulgar	Vulgar	Non-Abusive	Abusive
Bangla	906	1476	311	612	2081	2226	467	1954	739
Hindi	340	525	276	371	770	764	377	834	307
Gujarati	404	291	194	219	670	592	297	721	168
Bodo	227	151	0	39	339	271	107	381	77

Table 2
Statistics of Sentiment, Sarcasm, Vulgarity, and Abuse across Bangla, Hindi, Gujarati, and Bodo training datasets.

Abuse labels also display asymmetry. While Bangla and Hindi contain more non-abusive than abusive samples, the gap is large (Bangla: 1,954 vs. 739; Hindi: 834 vs. 307). Gujarati follows the same trend, whereas Bodo has a relatively lower but still imbalanced distribution (381 non-abusive vs. 77 abusive).

Overall, each of the four languages demonstrates clear class imbalance across sentiment, sarcasm, vulgarity, and abuse labels, which highlights the need for careful modeling strategies and balanced evaluation. Figure 3 provides a graphical visualization of these distributions.

5. Results

The macro F1-score computes the F1-score independently for each label: sentiment, abuse, vulgarity, sarcasm, then averages them. This metric places greater emphasis on minority classes, penalizing systems more heavily when performance is weak on underrepresented labels. The choice of an F1 variant depends on the task objectives and class distribution; however, due to the significant class imbalance typically present in hate speech detection, the macro F1-score is the most appropriate evaluation measure. For participant's system run submissions and evaluation in our task, we rely on the Kaggle platform. Figure 4 displays a screenshot of the Kaggle leaderboard used for run submissions. Separate Kaggle competition pages are provided for Bangla [3], Hindi [4], Gujarati [5] and Bodo [6] to enable participants to submit their experimental results.

Overall, 45 participants register for the task HASOC meme. In the Bangla task, 17 teams made 69 submissions, while 18 teams submitted 82 runs in the Hindi task, for Gujarati 15 teams made 68 submissions, and for the Bodo task, 15 teams submitted a total of 87 runs.

The performance of the best classification algorithms for Bangla, Hindi, Gujarati, and Bodo are Macro F1 measures of 0.6275, 0.6570, 0.6750, and 0.6312, respectively. The results for Bangla, Hindi, and Gujarati datasets are shown in Table 3, Table 4, Table 5, and Table 6, respectively.

[3] https://www.kaggle.com/competitions/hasoc-2025-meme-bangla (Access on 05.12.2025)
[4] https://www.kaggle.com/competitions/hasoc-2025-meme-hindi (Access on 05.12.2025)
[5] https://www.kaggle.com/competitions/hasoc-2025-meme-gujarati (Access on 05.12.2025)
[6] https://www.kaggle.com/competitions/hasoc-2025-meme-bodo (Access on 05.12.2025)

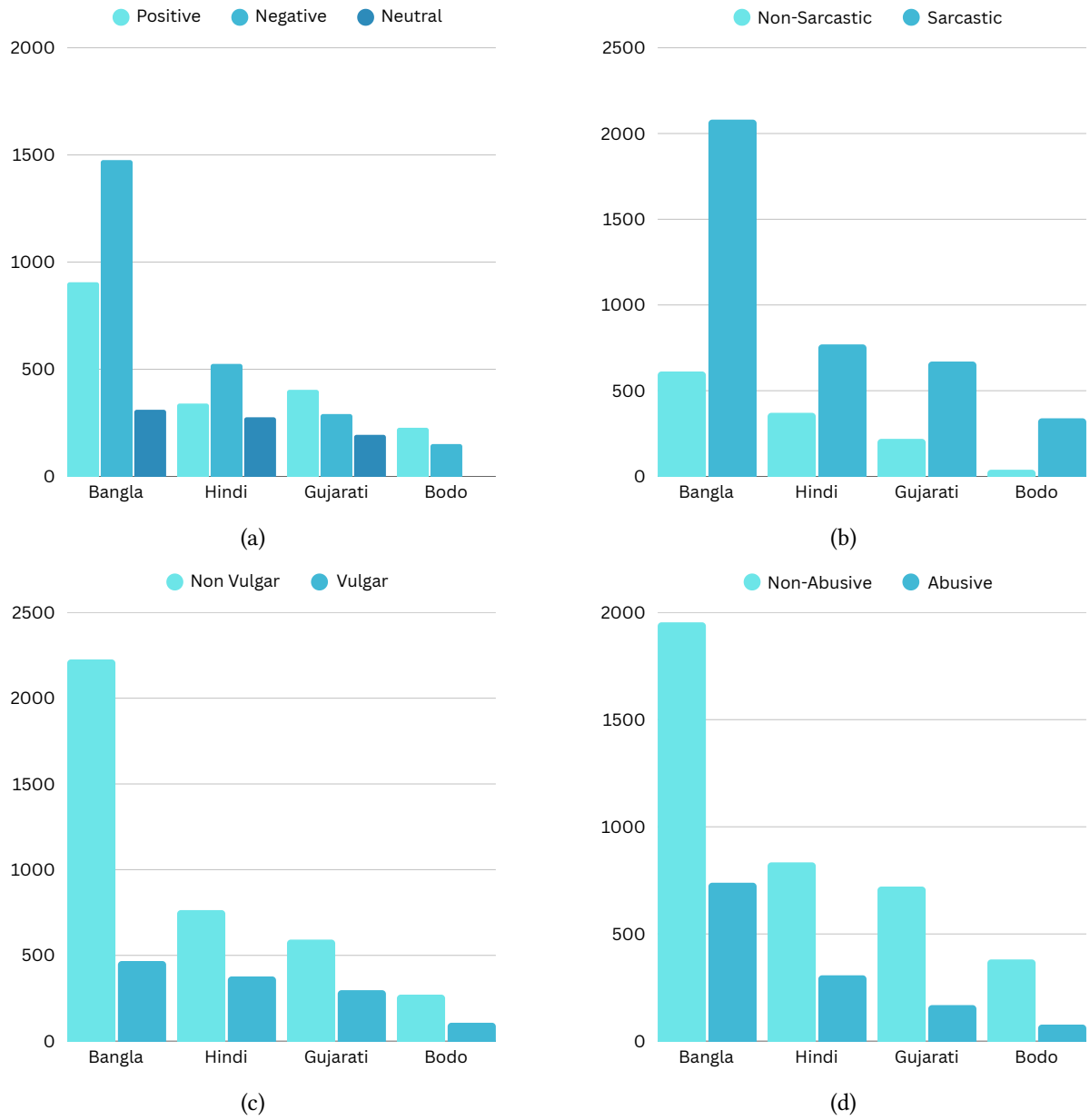

Figure 3: Graph visualization of Sentiment, Sarcasm, Vulgarity, and Abuse classes across Bangla, Hindi, Gujarati, and Bodo datasets.

6. Methodology

This section discusses the systems utilized by the participants.

- *FiRC-NLP* [32] explored three distinct approaches: prompt-based inference (Zero-Shot Classification and Few-Shot In-Context Learning via retrieval) using Gemini (Google's Gemini 2.5 Flash), cross-modality encoders that combine image and text, and a text-only modality leveraging Optical Character Recognition (OCR), OCR-based English translation, and image descriptions. Their final ensemble system fuses these modalities, demonstrating the effectiveness of multimodal fusion and robust training strategies.
- *CSIS BITS Pilani* [33] evaluated multiple dual-encoder architectures, combining the CLIP Vision Transformer with language-specific text models including MuRIL, XLM-Roberta, and M-BERT. Training is conducted using a 5-fold cross-validation strategy with a weighted loss function to counteract class imbalance, and performance is validated on the test set.

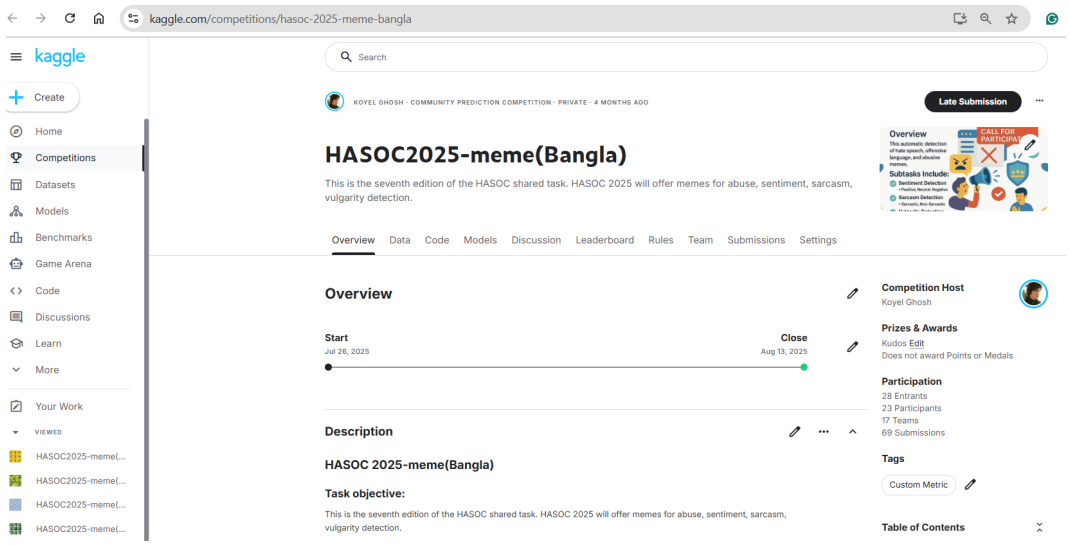

Figure 4: Screenshot of the HASOC-meme Kaggle website for run submission on Bangla (same for Hindi, Gujarati and Bodo).

Rank	Team	F1
1	FiRC-NLP [32]	0.62755
2	CSIS BITS Pilani [33]	0.61820
3	Golden Ratio [34]	0.61538
4	SCaLAR [35]	0.61034
5	NLPFusion [36]	0.60834
6	KK_NLP_AI_IIIT_Ranchi [37]	0.60595
7	Phinix (Abu Taher)	0.57563
8	IIT Dhanbad [38]	0.57209
9	DeepMeme	0.56203
10	CSE_SVNIT [39]	0.55476
11	MUCS [40]	0.53785
12	HASOC2025_meme (Baseline)	0.53185
13	YNU (Kongqiang Wang) [41]	0.52528
14	IReL [42]	0.50818
15	HASOC_2025 [43]	0.48746
16	VEL (Charmathi Rajkumar)	0.48331
17	DeepSemantics [44]	0.48211

Table 3
Results of the participants for the HASOC Meme 2025 (Bangla)

- *Golden Ratio* [34] integrated the CLIP vision encoder with tailored Transformer-based language models: MuRIL for Bangla, XLM-RoBERTa for Hindi, and Bodo. Textual and visual features are fused using a cross-attention mechanism to enhance classification accuracy.
- *SCaLAR* [35] experimented with two architectures: a CLIP-based model combining CLIP ViT visual features with multilingual Sentence Transformer embeddings, and a ViT + XLM-Roberta model where visual and textual features were concatenated for classification. The CLIP-based model achieved superior performance in Bangla and Gujarati, underscoring the effectiveness of contrastive pretraining for aligning visual and linguistic representations.
- *NLPFusion* [36] For Bangla memes, authors employed 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

Rank	Team	F1
1	FiRC-NLP [32]	0.65706
2	NLPFusion [36]	0.62398
3	KK_NLP_AI_IIIT_Ranchi [37]	0.59597
4	Golden Ratio [34]	0.59097
5	SCaLAR [35]	0.58468
6	IIT Dhanbad [38]	0.57417
7	CSE_SVNIT [39]	0.57198
8	CSIS BITS Pilani [33]	0.56788
9	VEL (Charmathi Rajkumar)	0.56769
10	DeepSemantics [44]	0.54989
11	HASOC2025_meme (Baseline)	0.54181
12	HASOC_2025 [43]	0.53602
13	FAST	0.52881
14	MUCS [40]	0.52497
15	YNU (Kongqiang Wang) [41]	0.51985
16	IReL [42]	0.46302
17	NITA_ICFAI	0.34037

Table 4
Results of the participants for the HASOC Meme 2025 (Hindi)

Rank	Team	F1
1	FiRC-NLP [32]	0.67501
2	NLPFusion [36]	0.63436
3	MUCS [40]	0.61848
4	SCaLAR [35]	0.61715
5	KK_NLP_AI_IIIT_Ranchi [37]	0.61409
6	CSIS BITS Pilani [33]	0.60018
7	IReL [42]	0.59196
8	IIT Dhanbad [38]	0.58879
9	CSE_SVNIT [39]	0.58221
10	YNU (Kongqiang Wang) [41]	0.56253
11	VEL (Charmathi Rajkumar)	0.54678
12	HASOC2025_meme (Baseline)	0.49293
13	DeepSemantics [44]	0.42035
14	HASOC_2025 [43]	0.34472

Table 5
Results of the participants for the HASOC Meme 2025 (Gujarati)

diversity and code-mixed content.

- *KK_NLP_AI_IIIT_Ranchi* [37] first experimented with text-only, low-resource models that relied solely on OCR-extracted text. They fine-tuned a multilingual IndicBERT backbone and observed strong performance for Hindi, Gujarati, and Bodo but weak results for Bangla. To investigate whether the poor Bangla performance was due to IndicBERT limitations, they then tested monolingual BERT-based models for each language. These included bengali-bert, hindi-bert-v2, gujarati-bert, and pretrained-bodo-legal-bert. However, the monolingual backbones did not yield substantial improvements, and in some cases performed worse than the multilingual model. Secondly, They incorporated the CLIP Vision model to combine image features with BERT-based text features using simple feature concatenation. They tested both multilingual and monolingual text backbones with CLIP, finding improvements for Bangla and Gujarati but little to no gain for Hindi and consistently poor results for Bodo. Because CLIP failed to extract meaningful cross-modal features for Indian languages, cross-attention fusion was discarded and only concatenation-based fusion was used. They then evaluated monolingual BERT backbones with CLIP, which further improved performance for Bangla, Gujarati, and Hindi. Finally, due to

Rank	Team	F1
1	NLPFusion [36]	0.63128
2	FiRC-NLP [32]	0.62217
3	CNLP-UPES [45]	0.60921
4	SCaLAR [35]	0.60393
5	CSIS BITS Pilani [33]	0.59969
6	CSE_SVNIT [39]	0.58730
7	IIT Dhanbad [38]	0.58186
8	HASOC_2025 [43]	0.57776
9	KK_NLP_AI_IIIT_Ranchi [37]	0.57184
10	Golden Ratio [34]	0.56202
11	DeepSemantics [44]	0.56040
12	MUCS [40]	0.55215
13	VEL (Charmathi Rajkumar)	0.54566
14	IReL [42]	0.50111
15	HASOC2025_meme (Baseline)	0.39221

Table 6
Results of the participants for the HASOC Meme 2025 (Bodo)

consistently weak results for Bodo with image features, they submitted a text-only IndicBERT-v2 model for that language.

- *IIT Dhanbad* [38] developed a multimodal deep learning framework for five subtasks—Sentiment, Sarcasm, Vulgarity, Abuse, and Target Communities—to automatically classify hateful and offensive memes. A task-specific strategy was used, where each label was handled by a dedicated multimodal model using various text encoders (XLM-R, mBERT, MuRIL, BanglaBERT) and image encoders (EfficientNet, DenseNet, VGG19, ResNet), along with techniques like gated attention and weighted losses. Extensive experiments showed that these task-specific pipelines performed best when combined through ensemble methods.
- *CSE_SVNIT* [39] leveraged pre-trained transformer models including XLM-ROBERTa, IndicBERT, MuRIL, and mBERT enhanced with convolutional neural network (CNN) layers in different Indian languages. Experimental results show that MuRIL provides optimal performance and outperforms other transformer models in low-resource Indian languages.
- *MUCS* [40] integrated transformer-based text encoders (Indic-BERT, MuRIL, XLM-Roberta) with convolutional and transformer-based vision models (ResNet, EfficientNet, ViT) using two fusion mechanisms -concatenation and attention-based strategies, to effectively capture the complementary cues from both modalities.
- *HASOC2025_meme (Baseline)* system employs zero-shot prompting using the Qwen2-7B-Instruct model [7]. Without any task-specific fine-tuning, the model is queried directly with carefully designed prompts to classify memes across the target categories. This baseline demonstrates the effectiveness of lightweight, instruction-tuned language models in handling multimodal meme understanding tasks, especially when annotated training data is limited.
- *YNU* [41] experimented with Gaussian Naive Bayes, Logistic Regression, K-Neighbors Classifier, Support Vector Machine, Decision Tree Classifier, Linear SVC, Random Forest Classifier, later use Ensemble Learning.
- *IReL* [42] developed and evaluated two independent runs under the team name IReL. Run 1 employed XLM-RoBERTa fine-tuned separately for each language (Bangla, Hindi, Gujarati, and Bodo), leveraging multilingual transformer embeddings to capture semantic nuances in noisy meme texts. Run 2 applied a zero-shot approach using ChatGPT specifically for Bodo, addressing the severe lack of annotated resources for this language. Both approaches processed text-based meme content exclusively.

[7] https://huggingface.co/Qwen/Qwen2-7B-Instruct (Access on 05.12.2025)

- *HASOC_2025* [43] extracted image features using ResNet-101 and combined them with OCR-based text features from the HASOC-2025 dataset to classify content as abusive/vulgar or not. They addressed class imbalance using ADASYN and trained separate XGBoost classifiers for each language. Their experiments showed that this multimodal approach works well across different scripts, even with OCR noise, demonstrating the effectiveness of combining deep image embeddings with traditional machine learning for multilingual hate speech detection.
- *DeepSemantics* [44] introduced a multimodal comparative analysis for detecting content that is, sarcastic, abusive, and vulgar. In addition to that, the sentiment of any content are also captured. The models are applied on the four different datasets consisting of different languages: Hindi, Bodo, Gujarati, and Bengali. Experiments are performed on various models, of which VisualBert was found to give the best performance. This model is most effective for detecting hateful and offensive content in multilingual datasets.
- *FAST* [46] used (i) a tailored preprocessing pipeline for noisy OCR and Hindi, English code-mixing using curated stopword and vulgar dictionaries, (ii) a combination of lightweight classical models (TF–IDF + Random Forest) with neural approaches (CNN, BiLSTM, ResNet50).
- *CNLP-UPES* [45] proposed 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.

7. Conclusion and Future Work

In this work, we presented the results of the HASOC Meme 2025 shared task on abusive meme identification across four languages: Bangla, Hindi, Gujarati, and Bodo. The task attracted a significant number of participants, reflecting the growing research interest in multimodal hate and abuse detection for Indian languages. Teams explored a wide range of approaches, including Zero-Shot Classification, Few-Shot In-Context Learning with retrieval-based prompting, and CLIP-based Vision Transformers, demonstrating the rapid evolution of multilingual and multimodal learning techniques.

While the shared task yielded promising insights, it also highlighted substantial gaps, especially for low-resource languages. Further experimentation is essential, particularly in building richer multimodal datasets, improving cross-lingual transfer, and developing robust models capable of handling linguistic diversity and culturally grounded abusive content. Continued efforts in these directions will enable more inclusive and effective solutions for combating harmful content across a broader spectrum of languages.

8. Acknowledgments

We thank Mr. Atri Chandra, Mr. Farhan Chowdhury, Mr. Animesh Howladar, Mr. Suvankar Dey, Mr. Trideep Ghosh, Mr. Alankrita Kumari, Mr. Ayush Lodh, Mr. Priyanka Das, Ms. Aveepsa Hatua, Mr. Bhaskar Pal, Mr. Gajjar Kandarp Dipakkumar, Ms. Vanpariya Palak Govindbhai, Mr. Patel Vraj Bipinbhai, Ms. Nijira Mushahary for the dataset annotation. We also thank the FIRE and HASOC organizers for their support in organizing the track. We thank all participants for their submissions and their valuable work.

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] R. Dawkins, The Selfish Gene, new edition ed., Oxford University Press, Oxford, UK, 1989.

[2] S. Pramanick, S. Sharma, D. Dimitrov, M. S. Akhtar, P. Nakov, T. Chakraborty, Momenta: A multimodal framework for detecting harmful memes and their targets, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 4439–4455.

[3] N. Statt, Youtube is facing a full-scale advertising boycott over hate speech (2017). URL: https://www.theverge.com/2017/3/24/15054878/youtube-advertising-boycott-hate-speech-extremist-content, accessed: [add access date].

[4] J. S. Vedeler, T. Olsen, J. Eriksen, Hate speech harms: A social justice discussion of disabled norwegians' experiences, Disability & Society 34 (2019) 368–383. doi:10.1080/09687599.2018.1436033.

[5] J. McSwiney, M. Vaughan, A. Heft, M. Hoffmann, Sharing the hate? Memes and transnationality in the far right's digital visual culture, Information, Communication & Society 24 (2021) 2502–2521. URL: https://www.tandfonline.com/doi/full/10.1080/1369118X.2021.1961006. doi:10.1080/1369118X.2021.1961006.

[6] U. K. Schmid, Humorous hate speech on social media: A mixed-methods investigation of users' perceptions and processing of hateful memes, New Media & Society (2023). URL: http://journals.sagepub.com/doi/10.1177/14614448231198169. doi:10.1177/14614448231198169.

[7] R. Gomez, J. Gibert, L. Gomez, D. Karatzas, Exploring hate speech detection in multimodal publications, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1470–1478.

[8] B. O. Sabat, C. Canton Ferrer, X. Giro-i Nieto, Hate speech in pixels: Detection of offensive memes towards automatic moderation, arXiv preprint arXiv:1910.02334 (2019). URL: https://arxiv.org/abs/1910.02334.

[9] S. Suryawanshi, B. R. Chakravarthi, M. Arcan, P. Buitelaar, Multimodal meme dataset (multioff) for identifying offensive content in image and text, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, 2020, pp. 32–41.

[10] S. Pramanick, D. Dimitrov, R. Mukherjee, S. Sharma, M. S. Akhtar, P. Nakov, T. Chakraborty, Detecting harmful memes and their targets, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 2783–2796.

[11] S. Pramanick, S. Sharma, D. Dimitrov, M. S. Akhtar, P. Nakov, T. Chakraborty, Momenta: A multimodal framework for detecting harmful memes and their targets, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 4439–4455.

[12] S. Modha, T. Mandl, P. Majumder, D. Patel, Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in Indo-European Languages, in: Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, Dec. 12-15, volume 2517, CEUR-WS.org, 2019, pp. 167–190. URL: http://ceur-ws.org/Vol-2517/T3-1.pdf.

[13] B. O. Sabat, C. C. Ferrer, X. G. i Nieto, Hate speech in pixels: Detection of offensive memes towards automatic moderation, 2019. URL: https://arxiv.org/abs/1910.02334. arXiv:1910.02334.

[14] R. Gomez, J. Gibert, L. Gomez, D. Karatzas, Exploring hate speech detection in multimodal publications, 2019. URL: https://arxiv.org/abs/1910.03814. arXiv:1910.03814.

[15] M. Chandra, D. Pailla, H. Bhatia, A. Sanchawala, M. Gupta, M. Shrivastava, P. Kumaraguru, "subverting the jewtocracy": Online antisemitism detection using multimodal deep learning, in: Proceedings of the 13th ACM Web Science Conference 2021, WebSci '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 148–157. URL: https://doi.org/10.1145/3447535.3462502. doi:10.1145/3447535.3462502.

[16] S. Suryawanshi, B. R. Chakravarthi, M. Arcan, P. Buitelaar, Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, ELRA, Marseille, France, 2020, pp. 32–41. URL: https://aclanthology.org/2020.trac-1.6.

[17] S. Pramanick, D. Dimitrov, R. Mukherjee, S. Sharma, M. S. Akhtar, P. Nakov, T. Chakraborty,

Detecting harmful memes and their targets, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 2783–2796. URL: https://aclanthology.org/2021.findings-acl.246/. doi:10.18653/v1/2021.findings-acl.246.

[18] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, C. A. Fitzpatrick, P. Bull, G. Lipstein, T. Nelli, R. Zhu, et al., The hateful memes challenge: competition report, in: NeurIPS 2020 Competition and Demonstration Track, PMLR, 2021, pp. 344–360. URL: https://proceedings.mlr.press/v133/kiela21a.html.

[19] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, SemEval-2022 Task 5: Multimedia automatic misogyny identification, in: Proceedings of the 16th International Workshop onSemantic Evaluation (SemEval), ACL, Ann Arbor, Michigan, 2022. doi:10.18653/v1/2022.semeval-1.74.

[20] L. Plaza, J. Carrillo-de Albornoz, E. Amigó, J. Gonzalo, R. Morante, P. Rosso, D. Spina, B. Chulvi, A. Maeso, V. Ruiz, Exist 2024: sexism identification in social networks and memes, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2024, pp. 498–504.

[21] D. S. M. Pandiani, E. T. K. Sang, D. Ceolin, Toxic memes: A survey of computational perspectives on the detection and explanation of meme toxicities, arXiv preprint arXiv:2406.07353 (2024).

[22] M. R. Karim, S. K. Dey, T. Islam, M. Shajalal, B. R. Chakravarthi, Multimodal hate speech detection from bengali memes and texts, 2022. URL: https://arxiv.org/abs/2204.10196. arXiv:2204.10196.

[23] E. Hossain, O. Sharif, M. M. Hoque, MUTE: A multimodal dataset for detecting hateful memes, in: Y. Hanqi, Y. Zonghan, S. Ruder, W. Xiaojun (Eds.), Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop, Association for Computational Linguistics, Online, 2022, pp. 32–39. URL: https://aclanthology.org/2022.aacl-srw.5/. doi:10.18653/v1/2022.aacl-srw.5.

[24] M. Das, A. Mukherjee, BanglaAbuseMeme: A dataset for Bengali abusive meme classification, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 15498–15512. URL: https://aclanthology.org/2023.emnlp-main.959/. doi:10.18653/v1/2023.emnlp-main.959.

[25] H. Kirk, Y. Jun, P. Rauba, G. Wachtel, R. Li, X. Bai, N. Broestl, M. Doff-Sotta, A. Shtedritski, Y. M. Asano, Memes in the wild: Assessing the generalizability of the hateful memes challenge dataset, in: A. Mostafazadeh Davani, D. Kiela, M. Lambert, B. Vidgen, V. Prabhakaran, Z. Waseem (Eds.), Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), ACL, Online, 2021, pp. 26–35. URL: https://aclanthology.org/2021.woah-1.4. doi:10.18653/v1/2021.woah-1.4.

[26] G. Kumari, D. Bandyopadhyay, A. Ekbal, Emoffmeme: identifying offensive memes by leveraging underlying emotions, Multimedia Tools Appl. 82 (2023) 45061–45096. URL: https://doi.org/10.1007/s11042-023-14807-1. doi:10.1007/s11042-023-14807-1.

[27] K. Rajput, R. Kapoor, K. Rai, P. Kaur, Hate me not: Detecting hate inducing memes in code switched languages, 2022. URL: https://arxiv.org/abs/2204.11356. arXiv:2204.11356.

[28] K. Maity, P. Jha, S. Saha, P. Bhattacharyya, A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 1739–1749. URL: https://doi.org/10.1145/3477495.3531925. doi:10.1145/3477495.3531925.

[29] P. Jha, K. Maity, R. Jain, A. Verma, S. Saha, P. Bhattacharyya, Meme-ingful analysis: Enhanced understanding of cyberbullying in memes through multimodal explanations, in: Y. Graham, M. Purver (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 930–943. URL: https://aclanthology.org/2024.eacl-long.56/. doi:10.18653/v1/2024.eacl-long.56.

[30] G. Kumari, A. Sinha, A. Ekbal, A. Chatterjee, V. B. N, Enhancing the fairness of offensive memes detection models by mitigating unintended political bias, J. Intell. Inf. Syst. 62 (2024) 735–763. URL: https://doi.org/10.1007/s10844-023-00834-9. doi:10.1007/s10844-023-00834-9.
[31] S. Suryawanshi, B. R. Chakravarthi, P. Verma, M. Arcan, J. P. McCrae, P. Buitelaar, A dataset for troll classification of TamilMemes, in: G. N. Jha, K. Bali, S. L., S. S. Agrawal, A. K. Ojha (Eds.), Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 7–13. URL: https://aclanthology.org/2020.wildre-1.2/.
[32] F. Hassan, M. S. Jahan, E. Migaev, F. Mtumbuka, Bridging modalities for hate speech detection in memes, in: Working Notes of FIRE 2025 - Forum for Information Retrieval Evaluation, CEUR, 2025.
[33] R. Bohra, Y. Sharma, A multi-modal ensemble approach for hate speech and offensive content detection in indic memes, in: Working Notes of FIRE 2025 - Forum for Information Retrieval Evaluation, CEUR, 2025.
[34] T. Paul, A. Jamatia, Hate speech and offensive content identification in memes in bangla, hindi, and bodo, in: Working Notes of FIRE 2025 - Forum for Information Retrieval Evaluation, CEUR, 2025.
[35] S. S. Nandam, A. K. Madasamy, Multimodal hate speech and offensive content classification in code-mixed indian language memes, in: Working Notes of FIRE 2025 - Forum for Information Retrieval Evaluation, CEUR, 2025.
[36] S. Coelho, A. Hegde, A. M Shetty, Nitte, Hasoc-meme: Enhancing hate speech recognition in bengali, hindi, gujarati, and bodo memes using multimodal multitask transformers, in: Working Notes of FIRE 2025 - Forum for Information Retrieval Evaluation, CEUR, 2025.
[37] U. Kedia, S. Prakash, K. Kumari, K. Prakash, T. Kumar, A transformer-based approach to multimodal hateful meme classification, in: Working Notes of FIRE 2025 - Forum for Information Retrieval Evaluation, CEUR, 2025.
[38] T. Kumari, A. Das, A. Sarvaiya, Hateful and offensive meme detection in multimodal memes dataset for indo-aryan languages, in: Working Notes of FIRE 2025 - Forum for Information Retrieval Evaluation, CEUR, 2025.
[39] S. S. Sahu, J. Damor, A transformer-based model for hate speech detection in low resource indian languages, in: Working Notes of FIRE 2025 - Forum for Information Retrieval Evaluation, CEUR, 2025.
[40] R. B N, S. H L, Towards safer social media: Multimodal hate speech detection in memes across diverse indian languages, in: Working Notes of FIRE 2025 - Forum for Information Retrieval Evaluation, CEUR, 2025.
[41] W. Kongqiang, T. Qingli, Hate speech and offensive content identification in memes in bangla, hindi and gujarati using auxiliary text supervised learning, in: Working Notes of FIRE 2025 - Forum for Information Retrieval Evaluation, CEUR, 2025.
[42] K. Tewari, S. Chanda, A. Namdeo, S. Pal, Savior: Sentiment, sarcasm, abuse, and vulgarity in online realities (memes), in: Working Notes of FIRE 2025 - Forum for Information Retrieval Evaluation, CEUR, 2025.
[43] S. Singh, G. Kumar, J. P. Singh, S. K. Rai, K. Goswami, Multilingual hate speech classification in memes using ocr-extracted text and visual features, in: Working Notes of FIRE 2025 - Forum for Information Retrieval Evaluation, CEUR, 2025.
[44] A. Malviya, A. Basak, S. Choudhury, Fusionguard: Visual-linguistic representations for multilingual harmful content detection, in: Working Notes of FIRE 2025 - Forum for Information Retrieval Evaluation, CEUR, 2025.
[45] P. Dadure, S. Ghosh, Bodo meme classification using multimodal fusion and oversampling in hasoc-2025, in: Working Notes of FIRE 2025 - Forum for Information Retrieval Evaluation, CEUR, 2025.
[46] M. Rafi, Fast-hasoc 2025: Multimodal and multilingual approaches for hate speech and offensive content detection in hindi memes, in: Working Notes of FIRE 2025 - Forum for Information

Retrieval Evaluation, CEUR, 2025.