SAViOR: Sentiment, Sarcasm, Abuse, and Vulgarity in Online Realities (Memes)

Krishna Tewari[1,*], Supriya Chanda[2], Aditya Namdeo[1] and Sukomal Pal[1]

[1]Indian Institute of Technology (BHU), Varanasi, INDIA
[2]Bennett University, Greater Noida, INDIA

Abstract

Meme-based communication on social media often encodes sentiment, sarcasm, vulgarity, and abuse in subtle and context-dependent ways, posing significant challenges for automated detection systems. These challenges are further amplified in low-resource Indian languages, where code-mixing, transliteration, and the scarcity of annotated data complicate modeling efforts. As part of the HASOC-meme shared task at FIRE 2025, we 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 and were evaluated on the official leaderboard using the macro-F1 metric. Our systems achieved macro-F1 scores of 0.5082 (Bangla), 0.4630 (Hindi), 0.5920 (Gujarati), and 0.5011 (Bodo), demonstrating the efficacy of multilingual transformers for high-resource languages and the viability of zero-shot large language models for severely underrepresented languages. These results highlight the strengths of transformer-based architectures in handling linguistic diversity while exposing limitations in handling subtle contextual cues in memes.

Keywords

Social Media, Meme classification, Sentiment Analysis, Sarcasm Detection, Abuse Detection, XLM-RoBERTa

1. Introduction

The analysis of online memes presents a challenging problem in natural language processing, as these short textual and visual artifacts often encode sentiment, sarcasm, vulgarity, and abuse in subtle, context-dependent ways. Let $M = \{m_1, m_2, \ldots, m_N\}$ denote a corpus of memes, where each meme m_i is represented by a textual component t_i written in one or more languages, potentially code-mixed or transliterated. The task of automated classification is then to assign each meme m_i a set of labels $L_i = \{s_i, c_i, v_i, a_i\}$, corresponding to sentiment, sarcasm, vulgarity, and abuse, respectively.

This problem is particularly challenging in the Indian social media ecosystem, where multilingualism and code-mixing are pervasive. Users frequently combine native scripts with Latin transliterations, resulting in text sequences $t_i = (w_1, w_2, \ldots, w_T)$ where w_j may belong to different languages and scripts. Consequently, effective classification requires models capable of capturing both cross-lingual semantic embeddings and nuanced contextual cues within noisy, informal text. Furthermore, the data is often highly imbalanced, with fewer examples for low-resource languages such as Bodo, and sparse instances of sarcasm or abuse compared to neutral or positive sentiment.

The HASOC-meme shared task at FIRE 2025 [1, 2] formalizes this setting for four low-resource Indian languages: Hindi, Bengali, Gujarati, and Bodo. For each meme $m_i \in M$, systems are evaluated on their ability to predict the label set L_i accurately. To address these challenges, we develop a hybrid framework: Run 1 fine-tunes the XLM-RoBERTa model separately for each language to leverage multilingual transformer embeddings, while Run 2 applies a zero-shot ChatGPT approach specifically for Bodo to mitigate resource scarcity.

Forum for Information Retrieval Evaluation, December 17-20, 2025, India
*Corresponding author.
krishnatewari.rs.cse24@itbhu.ac.in (K. Tewari); suplife24@gmail.com (S. Chanda); aditya.namdeo.met23@itbhu.ac.in (A. Namdeo); spal.cse@itbhu.ac.in (S. Pal)
0009-0005-6599-9956 (K. Tewari); 0000-0002-6344-8772 (S. Chanda); 0000-0001-8743-9830 (S. Pal)

Formally, given a meme m_i with text t_i, our model f_θ predicts:

$$f_\theta(t_i) = \hat{L}_i = \{\hat{s}_i, \hat{c}_i, \hat{v}_i, \hat{a}_i\},$$

where $\hat{L}_i$ is the predicted label set and θ denotes the model parameters.

The remainder of this paper is structured as follows: Section 2 reviews prior research; Section 3 describes the HASOC 2025 dataset; Section 4 details our proposed methodology; Section 5 presents experimental results; and Section 6 concludes with key findings and future directions.

2. Related Work

The analysis of online memes presents a multifaceted challenge in natural language processing (NLP), especially within the context of low-resource Indian languages. Memes often encode sentiment, sarcasm, vulgarity, and abuse in subtle, context-dependent ways. The short textual sequences, often code-mixed and transliterated, require models that can capture cross-lingual semantic representations along with nuanced contextual cues [3, 4].

Early research in sentiment and offensive language detection relied on traditional machine learning approaches with handcrafted features, such as n-grams, sentiment lexicons, and syntactic patterns [5, 6, 7]. While effective for longer, formal text, these approaches struggle with the brevity and noisiness of memes. To address these limitations, deep learning models, particularly recurrent and convolutional networks, were introduced [8, 9]. However, these methods often fail to generalize across languages, especially for low-resource languages with limited annotated data.

Transformer-based architectures have become the dominant paradigm for multilingual NLP tasks. Multilingual BERT (mBERT) [10] and XLM-RoBERTa [11] have been widely adopted for cross-lingual text classification, demonstrating the ability to transfer knowledge across languages without requiring large parallel corpora. Specifically for Indian languages, the IndicNLP Suite [3] provides preprocessing tools such as tokenizers and normalizers for code-mixed text, while MuRIL [4] offers pre-trained embeddings tuned for multiple Indian languages. Dowlagar and Mamidi [12] further explored the combination of XLM-RoBERTa with convolutional neural networks for sentence-level classification in technical domains, which is applicable to meme text analysis.

Shared tasks have played a key role in benchmarking and advancing research on offensive language and hate speech detection. The HASOC (Hate Speech and Offensive Content) shared tasks, initiated in 2019, provided multilingual datasets for Hindi, Bengali, and other Indian languages [13, 14, 15]. The tasks focused on the detection of hate speech, offensive content, and profanity in social media text, highlighting challenges such as data imbalance, code-mixing, and transliteratioN. The Dravidian-CodeMix shared task at FIRE 2021 further extended this line of research by providing datasets in Tamil, Malayalam, and Kannada, emphasizing offensive language detection in code-mixed social media content [16].

Understanding memes often requires multimodal analysis. The Hateful Memes Challenge by Facebook AI [17] introduced a dataset combining textual and visual content, encouraging the development of models capable of reasoning across modalities. Suryawanshi, Chakravarthi, and others [18] proposed a taxonomy and dataset for detecting opinion manipulation in troll memes, reinforcing the importance of contextual and multimodal cues in meme comprehension.

Low-resource languages, such as Bodo, pose a significant challenge due to the scarcity of annotated data. Zero-shot and few-shot learning approaches have emerged as viable solutions. Large language models such as ChatGPT, pre-trained on extensive multilingual corpora, can be leveraged in zero-shot settings to classify memes in low-resource languages [19]. This approach mitigates the need for task-specific fine-tuning and enables practical solutions for severely underrepresented languages.

Despite the progress of transformer-based methods, several challenges remain. Sarcasm, cultural references, and subtle forms of abuse are often context-dependent and difficult for models to detect accurately [20, 21]. Additionally, code-mixed sequences with transliteration and spelling variation complicate tokenization and embedding representation. Research has shown that combining multilingual embeddings with task-specific fine-tuning and cross-lingual transfer can improve performance [22, 4].

Benchmark-driven research underscores the importance of shared tasks in evaluating system performance. HASOC datasets [23], for instance, have demonstrated that fine-tuned transformer models significantly outperform traditional approaches on sentiment, sarcasm, and abusive language classification [13]. Studies by Chanda et al. [24, 25, 26, 27] illustrate that language-specific fine-tuning, combined with cross-lingual embeddings, is effective for medium and high-resource Indian languages, while zero-shot inference remains the most practical solution for languages with extremely limited annotated data. Recent research also explores hybrid approaches, integrating multimodal information and linguistic heuristics to improve robustness [18, 17, 28, 29].

In summary, prior work highlights several key themes: the necessity of multilingual and code-mixed text processing, the dominance of transformer-based architectures, the critical role of benchmark-driven evaluations such as HASOC and Dravidian-CodeMix, and the potential of zero-shot learning for low-resource languages. Building upon these strands, our work implements two independent runs for the HASOC-meme 2025 task: Run 1 fine-tunes XLM-RoBERTa separately for each language, while Run 2 applies zero-shot inference using ChatGPT for Bodo. This hybrid methodology leverages language-specific embeddings and zero-shot capabilities to address the linguistic diversity and annotation sparsity inherent in social media memes.

3. Dataset

The dataset provided in the HASOC-meme shared task at FIRE 2025 consisted of multimodal memes (image + text) in four Indian languages: Hindi, Bengali, Gujarati, and Bodo. Each meme was annotated for four text-based classification subtasks: Sentiment Detection, Sarcasm Detection, Vulgarity Detection, and Abuse Detection. An additional task, Target Community Identification, was initially included but later excluded from evaluation due to a high prevalence of null values and ambiguities in group assignments. Consequently, our experiments focused exclusively on the first four classification tasks. The detailed statistics of each dataset is presented in Table 1.

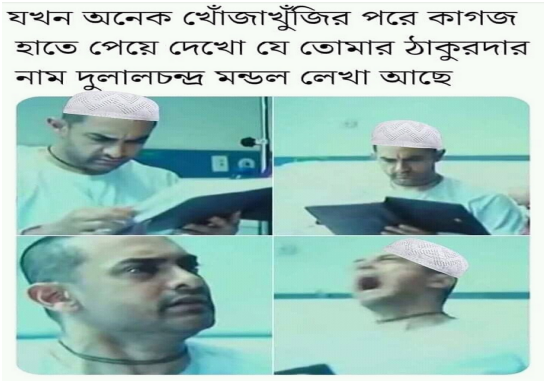
Table 1
Dataset distribution for HASOC 2025 tasks across Hindi, Bengali, Bodo, and Gujarati languages. The table shows the number of examples per class for Sentiment Analysis, Sarcasm Detection, Vulgarity Detection, and Abusive Content Classification.

Language	Sentiment			Sarcasm		Vulgarity		Abuse	
	Neg (-1)	Neu (0)	Pos (1)	Non-S (0)	S (1)	Non-V (0)	V (1)	Non-A (0)	A (1)
Hindi	525	276	340	371	770	764	377	834	307
Bengali	1476	311	906	612	2081	2226	467	1954	739
Bodo	151	0	227	39	339	269	107	301	77
Gujarati	291	194	404	219	670	592	297	721	168

The datasets exhibited significant class imbalance across all four subtasks, with minority classes (e.g., Vulgar or Abusive memes) being severely underrepresented. Exploratory analysis revealed several task-specific and target-group-specific trends. In the Gujarati dataset, abusive memes primarily targeted genders and individuals, whereas most vulgar memes were associated with gender. Sarcasm was frequent, but balanced sarcastic/non-sarcastic labels were observed mainly in memes targeting individuals. In the Hindi dataset, individual-targeted memes were mostly non-abusive and non-vulgar but often carried negative or neutral sentiment. Vulgar memes disproportionately targeted individuals and gender categories, while abusive memes were directed at individuals, gender, and "others" categories. Sarcasm was especially prevalent in memes targeting individuals.

Figures 1a and 1b show example memes from the dataset, including the OCR-extracted text and their corresponding annotations for the Bengali and religion-targeted categories, respectively.

(a) **OCR-extracted text: "অনুজীব এবং পটস দের বানানো কিছু খাদ্য: ল্যাকটোব্যাসিলাস, মৌমাছি, ইস্ট, ভাম-প্যান্টি". Sentiment: Negative; Sarcasm: Sarcastic; Vulgarity: Not Vulgar; Abuse: Abusive; Target: Political.**



(b) **OCR-extracted text: "যখন অনেক খোঁজাখুঁজি করে কাগজ হাতে পেয়ে দেখা যে তোমার ঠাকুরদার নাম দুলালচন্দ্র মন্ডল লেখা আছে". Sentiment: Neutral; Sarcasm: Sarcastic; Vulgarity: Not Vulgar; Abuse: Abusive; Target: Religion.**

Figure 1: Side-by-side example memes from the dataset, with OCR-extracted text and annotations included.

4. Methodology

We address the problem of robust multimodal meme classification in a multilingual setting. Formally, given a meme composed of an image I and OCR-extracted text t in one of the target languages (Hindi, Bengali, Bodo, or Gujarati), the goal is to predict class labels $y = \{y_1, y_2, y_3, y_4\}$ corresponding to Sentiment Analysis, Sarcasm Detection, Vulgarity Detection, and Abusive Content Classification.

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} P(y \mid I, t),$$

where $\mathcal{Y}$ denotes the set of all possible label combinations across tasks.

Our solution leverages two stages: preprocessing, fine-tuning XLM-RoBERTa (XLM-R) on all languages (submitted as Run 1), and/or zero-shot classification using ChatGPT (submitted as Run 2) exclusively for the Bodo dataset.

4.1. Preprocessing

In the preprocessing phase, we performed several essential cleaning steps to prepare the dataset. Hashtags, punctuation marks, URLs, numbers, and user mentions were removed to eliminate noisy tokens that did not contribute meaningfully to the classification tasks. Emojis were converted into their corresponding text representations to preserve sentiment and emotion information in textual form. Extra spaces were eliminated to ensure clean and consistent input formatting.

An important step involved normalizing code-mixed content written in Latin script into the appropriate native script (Hindi, Bengali, Bodo, or Gujarati) using the Indic-NLP transliteration tool. This was critical for handling frequent use of Romanized text in social media posts and enabling the downstream models to process language-specific tokens effectively.

4.2. XLM-RoBERTa Fine-Tuning Across All Languages (Run 1)

To effectively model the multilingual and code-mixed text, we employed XLM-RoBERTa (XLM-R) [11], a transformer-based model pretrained on massive multilingual corpora. For each of the four classification tasks, the model predicts a label $\hat{y}_i$ given the OCR-extracted text t as input. The fine-tuning objective was formulated as a multi-class classification problem, optimized using cross-entropy loss.

We fine-tuned XLM-R separately for each language, utilizing task-specific training datasets. A low learning rate of 5×10^{-6} was adopted to stabilize training dynamics, accompanied by a cosine learning rate scheduler. The models were trained for 15–20 epochs with early stopping based on the validation macro-F1 score, which balances performance across all classes and mitigates bias towards majority labels.

To address the severe class imbalance, we weighted the loss function inversely proportional to class frequencies. Additionally, data oversampling techniques were applied to minority classes, particularly for the Vulgarity task which follows a tri-class scheme: Negative (-1), Neutral (0), and Positive (1). The final model used the `xlm-roberta-base` checkpoint from the Hugging Face library, with the input consisting solely of the tokenized OCR text. Empirical results demonstrated that XLM-R consistently outperformed other baselines, capturing the semantic subtleties of code-mixed and noisy text, and was selected as our official Run 1 submission.

4.3. Zero-Shot Classification with ChatGPT (Run 2)

Recognizing the extremely limited training data available for the Bodo dataset, we explored zero-shot classification using ChatGPT (GPT-4o-mini) to evaluate its robustness in low-resource settings. Unlike the fine-tuning approach, no task-specific training was conducted.

For each example, we constructed detailed prompts incorporating both the OCR-extracted text and the image modality. The prompt instructed the model to classify the meme strictly based on the given image and text into Sentiment, Sarcasm, Vulgarity, and Abuse categories. An example of the prompt template used is as follows:

Prompt = "Classify the following meme into Sentiment, Sarcasm, Vulgarity, and Abuse categories based on the image and OCR-extracted text."
Output format:
Sentiment: [Negative/Neutral/Positive],
Sarcasm: [Sarcastic/Non-Sarcastic],
Vulgarity: [Vulgar/Non-Vulgar],
Abuse: [Abusive/Non-Abusive]."

This approach enabled the model to generate fluent, contextually complete classification outputs without fine-tuning. Prompts were submitted in batches of 50 examples via the OpenAI API. The raw outputs were parsed to extract structured labels, discarding any extraneous commentary. The inclusion of the image alongside text proved essential for improving classification accuracy, particularly given the subtle and ambiguous nature of the content.

Overall, this zero-shot paradigm served as a complementary methodology to the XLM-R-based supervised fine-tuning, providing a practical solution for extremely low-resource scenarios.

5. Results

This section reports the official leaderboard results of the HASOC 2025 meme classification task across four languages: Bangla, Hindi, Gujarati, and Bodo.

5.1. Bangla

Table 2 presents the Bangla results. FiRC-NLP (0.6276), CSIS BITS Pilani (0.6182), and Golden Ratio (0.6154) dominated the top of the leaderboard, with SCaLAR and NLPFusion also close behind. IReL achieved a score of **0.5082**, ranking 14th and falling just below the baseline (0.5319). This outcome suggests that while IReL captured some discriminative features, its model struggled with Bangla-specific

orthography and cultural humor patterns. The below-baseline placement indicates potential limitations in script handling and adaptation to the semantic richness of Bangla memes.

5.2. Hindi

Table 3 shows the Hindi leaderboard. FiRC-NLP achieved the top score (0.6571), followed by NLPFusion (0.6240) and KK_NLP_AI_IIIT_Ranchi (0.5960). Many teams clustered between 0.57–0.59, underscoring the strength of the competition. IReL scored **0.4630**, ranking 16th and below the baseline (0.5418). This gap is larger than in Bangla, reflecting the greater difficulty of handling Hindi memes, which often involve code-mixing, transliteration, and sarcasm. The results indicate that IReL's system did not fully capture the complexity of Hindi's multilingual meme content.

Table 2
Leaderboard for Bangla.

Rank	Team	Score
1	FiRC-NLP	0.62755
2	CSIS BITS Pilani	0.61820
3	Golden Ratio	0.61538
4	SCaLAR	0.61034
5	NLPFusion	0.60834
6	KK_NLP_AI_IIIT_Ranchi	0.60595
7	Phinix (Abu Taher)	0.57563
8	IIT Dhanbad	0.57209
9	DeepMeme	0.56203
10	CSE_SVNIT	0.55476
11	MUCS	0.53785
12	HASOC2025_meme (Baseline)	0.53185
13	YNU (kongqiang wang)	0.52528
14	**IReL**	**0.50818**
15	HASOC_2025	0.48746
16	VEL (Charmathi Rajkumar)	0.48331
17	DeepSemantics	0.48211

Table 3
Leaderboard for Hindi.

Rank	Team	Score
1	FiRC-NLP	0.65706
2	NLPFusion	0.62398
3	KK_NLP_AI_IIIT_Ranchi	0.59597
4	Golden Ratio	0.59097
5	SCaLAR	0.58468
6	IIT Dhanbad	0.57417
7	CSE_SVNIT	0.57198
8	CSIS BITS Pilani	0.56788
9	VEL (Charmathi Rajkumar)	0.56769
10	DeepSemantics	0.54989
11	HASOC2025_meme (Baseline)	0.54181
12	HASOC_2025	0.53602
13	FAST	0.52881
14	MUCS	0.52497
15	YNU (kongqiang wang)	0.51985
16	**IReL**	**0.46302**
17	NITA_ICFAI	0.34037

5.3. Gujarati

Table 4 highlights the Gujarati results. FiRC-NLP (0.6750) and NLPFusion (0.6344) were the clear leaders, followed by MUCS, SCaLAR, and KK_NLP_AI_IIIT_Ranchi around 0.61. IReL achieved its best performance here, scoring **0.5920** and ranking 7th, well above the baseline (0.4929). This result demonstrates that IReL's model was able to align well with Gujarati meme characteristics, likely benefiting from more effective script handling and better adaptation. The placement in the upper half of the leaderboard shows competitiveness, though a gap remains compared to the leaders.

5.4. Bodo

Table 5 summarizes the Bodo results. NLPFusion (0.6313) narrowly surpassed FiRC-NLP (0.6222), with CNLP-UPES (0.6092) and SCaLAR (0.6039) also above 0.60. IReL obtained a score of **0.5011**, ranking 14th but importantly outperforming the baseline (0.3922). This indicates some generalization capability in a low-resource setting, though the gap to the top-performing systems remains considerable. The result reflects that IReL can adapt to Bodo, but stronger multilingual transfer techniques are needed to close the performance gap.

Table 4
Leaderboard for Gujarati.

Rank	Team	Score
1	FiRC-NLP	0.67501
2	NLPFusion	0.63436
3	MUCS	0.61848
4	SCaLAR	0.61715
5	KK_NLP_AI_IIIT_Ranchi	0.61409
6	CSIS BITS Pilani	0.60018
7	**IReL**	**0.59196**
8	IIT Dhanbad	0.58879
9	CSE_SVNIT	0.58221
10	YNU (kongqiang wang)	0.56253
11	VEL (Charmathi Rajkumar)	0.54678
12	HASOC2025_meme (Baseline)	0.49293
13	DeepSemantics	0.42035
14	HASOC_2025	0.34472

Table 5
Leaderboard for Bodo.

Rank	Team	Score
1	NLPFusion	0.63128
2	FiRC-NLP	0.62217
3	CNLP-UPES (Pankaj Dadure)	0.60921
4	SCaLAR	0.60393
5	CSIS BITS Pilani	0.59969
6	CSE_SVNIT	0.58730
7	IIT Dhanbad	0.58186
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 (Charmathi Rajkumar)	0.54566
14	**IReL**	**0.50111**
15	HASOC2025_meme (Baseline)	0.39221

5.5. Cross-Language Observations

Across the four languages, IReL's results exhibit marked variability. The strongest performance was in Gujarati (7th place, 0.5920), highlighting effective script handling and feature capture. In contrast, Hindi (0.4630) and Bangla (0.5082) results fell below the baseline, primarily due to transliteration inconsistencies, cultural humor cues, and label imbalance, which limited discriminative learning. Manual inspection suggests that many errors originated from sarcasm-sentiment confusion and Romanized abusive slang that XLM-R tokenization could not interpret. These findings motivate incorporating Indian-language specific pretraining (e.g., MuRIL, IndicBERT v2) and data augmentation through back-translation and code-mixing simulation to mitigate these shortcomings. Furthermore, we plan to conduct granular error analyses using interpretability tools (e.g., SHAP, LIME) to systematically diagnose class-wise failures.

Notably, the zero-shot ChatGPT run for Bodo achieved a macro-F1 of 0.5011, exceeding the baseline (0.3922) and validating large language models' viability in extremely low-resource settings.

6. Conclusion and Future Work

The HASOC 2025 meme shared task underscored the complexity of detecting sentiment, sarcasm, vulgarity, and abuse across Bangla, Hindi, Gujarati, and Bodo in highly code-mixed and noisy settings. IReL submitted two complementary runs: the first fine-tuned XLM-RoBERTa separately for each language, achieving macro-F1 scores of 0.5082 (Bangla), 0.4630 (Hindi), 0.5920 (Gujarati), and 0.5011 (Bodo), thereby demonstrating competitive performance in Gujarati but below-baseline outcomes in Bangla and Hindi where transliteration and cultural nuance posed challenges. The second run leveraged zero-shot prompting with ChatGPT for Bodo, surpassing the official baseline and illustrating the potential of large language models for severely under-resourced languages. Overall, these results highlight the strengths of multilingual transformers for medium and high-resource contexts, the utility of zero-shot inference in low-resource scenarios, and the persistent difficulty of handling sarcasm, transliterated slang, and community-specific references. These findings motivate several improvements: (i) diversifying model architectures beyond XLM-R to include MuRIL, IndicBERT, and distilled variants; (ii) enriching datasets through back-translation, balanced resampling, and multimodal fusion; (iii) implementing explainable-AI diagnostics to interpret misclassifications; and (iv) enhancing scalability through student–teacher distillation and quantized inference for real-world deployment. Collectively, these directions aim to produce robust, efficient, and interpretable meme-understanding systems that

scale across diverse Indic languages.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] 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.

[2] 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.

[3] D. Kakwani, A. Kunchukuttan, S. Golla, N. C. Gokul, A. Bhattacharyya, M. M. Khapra, P. Kumar, Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 4948–4961. doi:10.18653/v1/2020.findings-emnlp.445.

[4] S. Khanuja, D. Bansal, S. Mehtani, S. Khosla, A. Dey, B. Gopalan, D. K. Margam, P. Aggarwal, R. T. Nagipogu, S. Dave, S. Gupta, S. C. B. Gali, V. Subramanian, P. Talukdar, Muril: Multilingual representations for indian languages, 2021. arXiv:2103.10730, submitted 19 Mar 2021; revised 2 Apr 2021.

[5] B. Pang, L. Lee, Opinion mining and sentiment analysis, Foundations and Trends® in Information Retrieval 2 (2008) 1–135. doi:10.1561/1500000011.

[6] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Computational Linguistics 37 (2011) 267–307. URL: https://aclanthology.org/J11-2001/. doi:10.1162/COLI_a_00049.

[7] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, C. Potts, Recursive deep models for semantic compositionality over a sentiment treebank, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2013.

[8] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014.

[9] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: Proceedings of NAACL-HLT, 2016.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019.

[11] A. Conneau, G. Lample, Unsupervised cross-lingual representation learning at scale, in: Proceedings of ACL, 2020.

[12] S. Dowlagar, R. Mamidi, Multilingual pre-trained transformers and convolutional nn classification models for technical domain identification, in: Preprint, 2021. ArXiv:2101.00000.

[13] T. Mandl, S. Modha, G. K. Shahi, A. K. Jaiswal, D. Nandini, D. Patel, P. Majumder, J. Schäfer, Overview of the hasoc track at fire 2020: Hate speech and offensive content identification in indo-european languages, in: Working Notes of FIRE 2020 – Forum for Information Retrieval

Evaluation, volume 2826 of *CEUR-WS.org*, CEUR Workshop Proceedings, Hyderabad, India, 2020, pp. 87–111. doi:10.1145/3441501.3441517.

[14] 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.

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

[16] B. R. Chakravarthi, P. K. Kumaresan, R. Sakuntharaj, A. K. Madasamy, S. Thavareesan, B. Premjith, K. Sreelakshmi, S. C. Navaneethakrishnan, J. P. McCrae, T. Mandl, Overview of the hasoc-dravidiancodemix shared task on offensive language detection in tamil and malayalam, in: Working Notes of FIRE 2021 – Forum for Information Retrieval Evaluation, volume 3159 of *CEUR-WS.org*, CEUR Workshop Proceedings, Hyderabad, India, 2021. Track on offensive language identification in code-mixed Dravidian languages.

[17] D. Kiela, D. Hazarika, E. Fini, S. K. Deepak, S. Poria, The hateful memes challenge: Detecting hate speech in multimodal memes, in: arXiv preprint arXiv:2005.04790, 2020.

[18] S. Suryawanshi, B. R. Chakravarthi, M. Arcan, P. Buitelaar, Trollswithopinion: A taxonomy and dataset for predicting domain-specific opinion manipulation in troll memes, Multimedia Tools Appl. 82 (2022) 9137–9171. URL: https://doi.org/10.1007/s11042-022-13796-x. doi:10.1007/s11042-022-13796-x.

[19] Y. Bang, S. Karthik, R. Arora, Multilingual hate speech detection with limited resources, in: Proceedings of the 4th Workshop on NLP for Low-Resource Languages, ACL, 2023.

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

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

[22] S. Modha, T. Mandl, G. K. Shahi, H. Madhu, S. Satapara, T. Ranasinghe, M. Zampieri, Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages and conversational hate speech, in: Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation (FIRE '21), ACM, India, 2022. doi:10.1145/3503162.3503176.

[23] K. Ghosh, S. Saha, T. Mandl, S. Modha, Findings from shared tasks on hate speech detection: Performance patterns for low-resource languages, Pattern Recognition Letters (2025). URL: https://www.sciencedirect.com/science/article/pii/S0167865525003150. doi:https://doi.org/10.1016/j.patrec.2025.09.004.

[24] S. Chanda, A. Mishra, S. Pal, Advancing language identification in code-mixed tulu texts: Harnessing deep learning techniques, in: K. Ghosh, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2023 – Forum for Information Retrieval Evaluation (FIRE-WN 2023), volume 3681 of *CEUR Workshop Proceedings*, CEUR-WS.org, Goa, India, 2023. CC BY 4.0.

[25] S. Chanda, S. Ujjwal, S. Das, S. Pal, Fine-tuning pre-trained transformer based model for hate speech and offensive content identification in english indo-aryan and code-mixed (english-hindi) languages, in: Forum for Information Retrieval Evaluation (FIRE 2021) Working Notes, Forum for Information Retrieval Evaluation, 2021. Shared Task HASOC 2021.

[26] S. Chanda, S. D. Sheth, S. Pal, Coarse and fine-grained conversational hate speech and offensive

content identification in code-mixed languages using fine-tuned multilingual embedding, in: Forum for Information Retrieval Evaluation (FIRE 2022) Working Notes, Forum for Information Retrieval Evaluation, 2022. Shared Task HASOC 2022.

[27] A. Saroj, S. Chanda, S. Pal, Irlab iitv at semeval-2020 task 12: Multilingual offensive language identification in social media using svm, in: A. Hébelot, X. Zhu, A. Palmer, N. Schneider, J. May, E. Shutova (Eds.), Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval 2020), International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 2012–2016. doi:10.18653/v1/2020.semeval-1.265.

[28] K. Ghosh, N. K. Singh, J. Mahapatra, et al., SafeSpeech: a three-module pipeline for hate intensity mitigation of social media texts in Indic languages, Social Network Analysis and Mining 14 (2024). URL: https://doi.org/10.1007/s13278-024-01393-9. doi:10.1007/s13278-024-01393-9.

[29] 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.