FAST-HASOC 2025: Multimodal and Multilingual Approaches for Hate Speech and Offensive Content Detection in Hindi Memes

Muhammad Rafi[1], Saif Ur Rehman Awan[1], Ramsha Jat[1], Aiman Falak[1], Fatimah Ansari[1], Ahmed Raza[1] and Sagar Chabbriya[1]

[1]*FAST National University of Computer and Emerging Sciences, Karachi, Pakistan*

Abstract

This paper describes our Team FAST's participation in the HASOC 2025 shared task on offensive content detection in Hindi memes. Our system achieved an overall rank of 13 with a score of 0.52881. The dataset consists of multimodal samples with OCR-extracted text and raw images, annotated across four subtasks: sentiment, sarcasm, vulgarity, and abuse detection. We propose a multimodal framework that integrates classical machine learning and deep learning models. Our contributions are as follows: (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). Code, data splits, and preprocessing resources are available at: https://github.com/fatimahansari/hindi-HASOC-2025.

Keywords

Hate Speech Detection, Multimodal NLP, Hindi Memes, TF–IDF, Random Forest, CNN, BiLSTM, ResNet50, HASOC 2025

1. Introduction

Offensive content on social media spans textual and visual modalities, often with subtle cues, code-mixing, and transliteration. The HASOC 2025 shared task [1] focuses on Hindi memes, combining OCR-extracted text and meme images, with subtasks in sentiment, sarcasm, vulgarity, and abuse detection. These challenges are amplified by noisy OCR, informal Hinglish, and implicit insults. We present our approach, which combines both classical and neural models under a multimodal pipeline. Unlike prior editions focusing on monomodal text, our system explicitly fuses image and text features for vulgarity and incorporates Hindi-specific lexicons for preprocessing. Our contributions are:

- A preprocessing framework for noisy OCR and Hindi–English code-mixing.
- A multimodal architecture combining TF-IDF + Random Forest with ResNet50.

2. Related Work

Hate speech and offensive content detection has been widely studied in the context of Indic languages, particularly through the HASOC shared tasks [2, 3, 4]. Earlier editions primarily focused on monolingual or text-based systems, while HASOC 2025 introduced multimodal Hindi memes, making the task more challenging.

Multimodal abusive meme classification gained growing interest with datasets like BanglaAbuseMeme [5], showing that images carry contextual cues that textual models alone may miss.

Hindi and code-mixed abusive language remains difficult due to transliteration, lexical variation, and implicit insults. Data augmentation and lexicon-aware approaches have shown improvements

Forum for Information Retrieval Evaluation (FIRE), December 17–20, 2025, India

muhammad.rafi@nu.edu.pk (M. Rafi); saifurrehman@nu.edu.pk (S. U. R. Awan); ramsha.jat@nu.edu.pk (R. Jat); aiman.falak@nu.edu.pk (A. Falak); fatimahansari614@gmail.com (F. Ansari); ahmedraza9332@gmail.com (A. Raza); sagarchhabriya34@gmail.com (S. Chabbriya)

0000-0002-3673-5979 (M. Rafi); 0009-0004-4235-1244 (S. U. R. Awan)

[6, 7, 8]. Transformer architectures such as BERT [9] and image transformers [10] have achieved strong performance in offensivecontent detection, and pipelines for safer text moderation continue to evolve [11]. However, these models require large annotated datasets and high compute, which limits their effectiveness in noisy, low-resource scenarios like HASOC 2025. Our work differs by focusing on lightweight yet robust multimodal methods, integrating lexicon-informed preprocessing with classical and neural models, optimized for noisy OCR data.

3. Dataset and Resources

We used the HASOC 2025 Hindi meme dataset [1]:

- **Train:** 1133 samples with labels.
- **Test:** 767 samples.

Each sample includes an image and OCR text annotated for four subtasks. Additional resources:

- `hindi_stopwords.json` — curated Hindi/Hinglish stopword list.
- `hindi-offensive-words-original.json` — offensive lexicon mapped to neutral terms.

These resources are publicly released with our code repository.

4. System Architecture

Figure 1 shows the system pipeline: preprocessed text is passed through task specific models; images are processed for vulgarity detection and fused with text predictions.

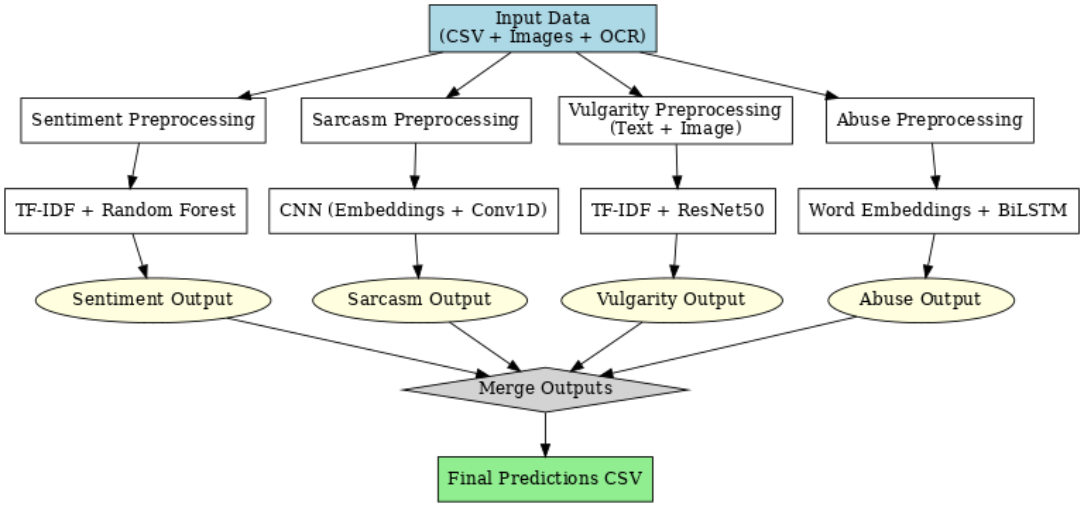

Figure 1: Overall multimodal system pipeline for HASOC 2025 subtasks.

4.1. Preprocessing

Our text preprocessing includes:

1. Cleaning URLs, emails, non-Devanagari symbols.
2. Stopword removal (Hindi/Hinglish).
3. Offensive word replacement using the vulgar dictionary.
4. Tokenization and language-aware normalization. These resources are publicly released with our code [12, 13].

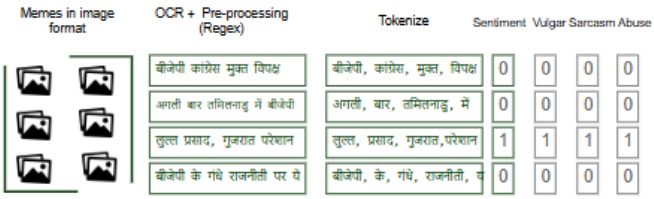

Figure 2: Pre-processing the OCR text

4.2. Models

Sentiment: For sentiment detection, we chose a combination of TF–IDF features and a Random Forest classifier. TF–IDF is effective in representing textual data from short social media posts and OCR text because it captures the importance of words and bigrams while ignoring overly frequent stopwords. Random Forest, being an ensemble of decision trees, provides robustness against noisy data and works well with sparse, high-dimensional inputs. This model was selected because sentiment cues in Hindi memes are often expressed through explicit keywords or short phrases, making classical feature-based approaches suitable. Additionally, Random Forests handle class imbalance relatively well and offer interpretability compared to deep models.

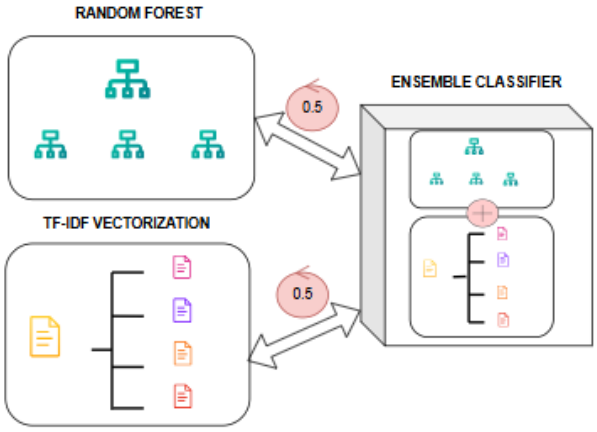

Figure 3: Sentiment detection model using TF–IDF vectorization and Random Forest.

Sarcasm: Sarcasm is typically expressed through subtle lexical patterns, wordplay, and local context within a short sequence of text. To model this, we employed a Convolutional Neural Network (CNN) with an embedding layer, 1D convolutional filters, and global max pooling. The CNN captures n-gram level features by sliding filters over word embeddings, allowing it to learn important combinations of words that signal sarcasm. This architecture is lightweight compared to transformers but effective for short-text sarcasm detection, which often relies on key sarcastic cues rather than long-range dependencies. We chose CNNs because they generalize well on small datasets, train faster, and are less prone to overfitting than more complex models.

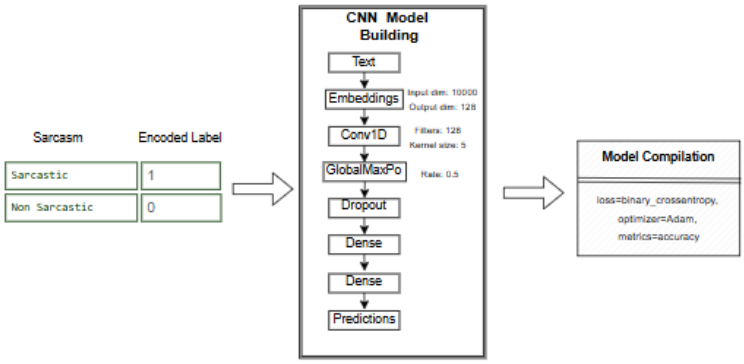

Figure 4: Sarcasm detection model using CNNs.

Vulgarity: This multimodal setup allows the system to leverage complementary strengths, while textual features capture linguistic vulgarity, image embeddings contribute crucial context when offensiveness is implied visually rather than verbally. The decision level fusion also provides robustness, as errors in one modality can be compensated by the other, leading to more stable and consistent predictions across a wide variety of meme formats. Furthermore, the use of a relatively lightweight architecture like ResNet50 ensures faster inference and reduced computational overhead, making the approach more practical for real world deployment where large volumes of memes need to be processed efficiently. This balance between accuracy, efficiency, and adaptability is particularly important for social media platforms, where offensive content spreads rapidly, and automated systems must detect problematic material in near real-time while maintaining scalability across millions of daily uploads.

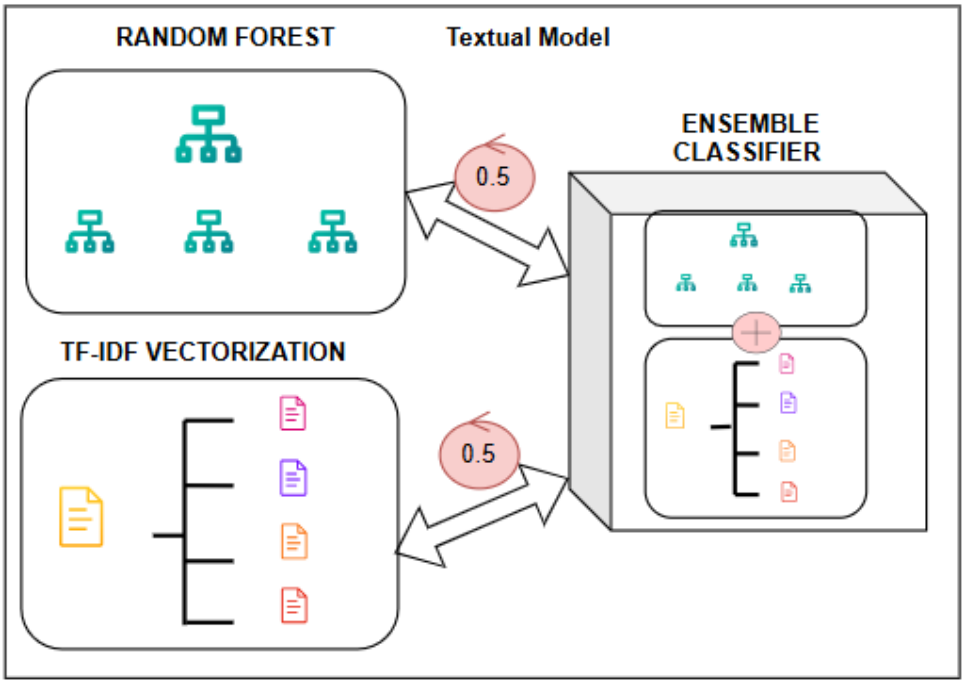

Figure 5: Multimodal vulgarity detection model combining text (TF-IDF + RF) features.

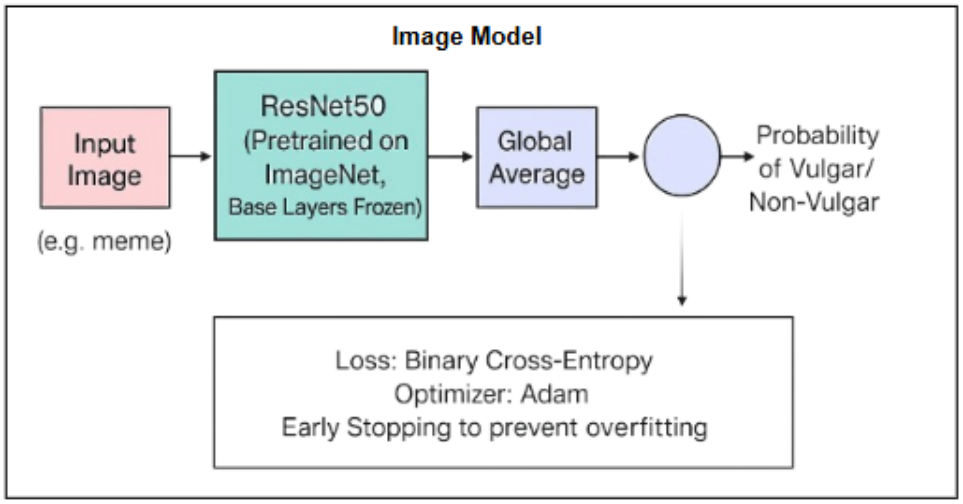

Figure 6: Multimodal vulgarity detection model combining image (ResNet50) features.

Abuse: Abuse detection is more complex than sentiment or vulgarity because abusive language is often indirect, context-dependent, and highly code-mixed, especially in spaces where users switch between Hindi and English. To capture these sequential cues, we used a Bidirectional LSTM that processes text in both directions, allowing it to model long-range dependencies and subtle indicators that simpler models miss. The model was initialized with Hindi FastText embeddings, which provide strong semantic coverage even for rare or morphologically rich words. We also added a binary lexicon feature to directly flag explicit slurs, while the text itself was normalized with placeholder replacements to ensure the model focused on context. This hybrid setup balances contextual understanding with explicit-term detection, making it well-suited for noisy social media text where abuse can be both overt and implicit.

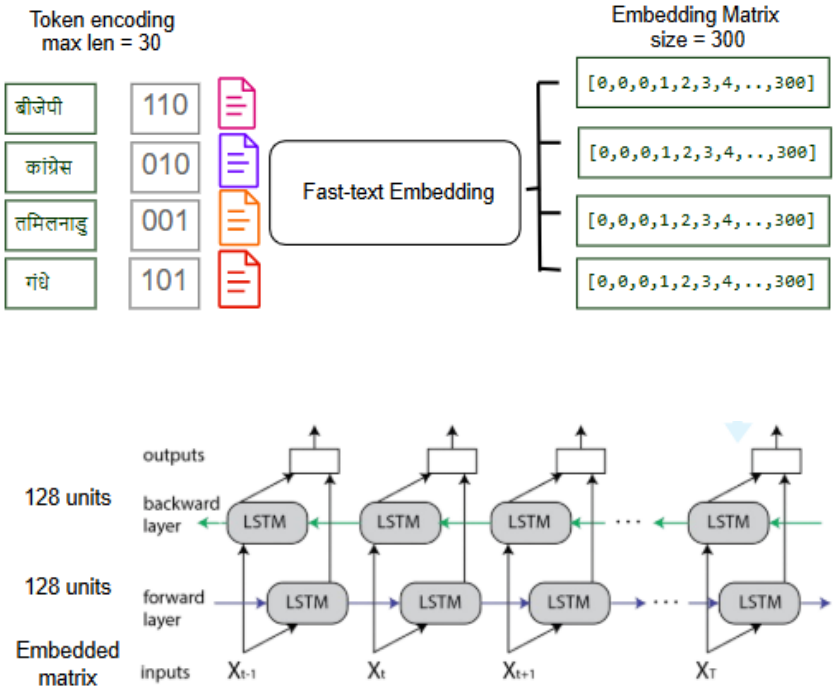

Figure 7: BiLSTM model for abuse detection with FastText embeddings and lexicon features.

Why not Transformers? Although transformer-based models such as BERT [9] and mBERT have shown strong results in offensive language detection, we deliberately chose not to rely on them as our primary models in this work. There are three key reasons. First, the HASOC 2025 dataset for Hindi memes is relatively small (just over 1100 training examples), which makes fine-tuning large transformer models prone to overfitting. In contrast, lighter models such as TF–IDF + Random Forest or CNNs are more data-efficient and generalize better in low-resource settings. Second, OCR-extracted Hindi text is noisy and often code-mixed, with Romanized tokens that pretrained multilingual transformers do not handle well without extensive normalization. Classical models and BiLSTMs with FastText embeddings proved more robust under these conditions, especially when augmented with curated lexicons. Finally, computational efficiency was an important consideration: Random Forests, CNNs, and BiLSTMs are significantly faster to train and deploy, making them practical for iterative experimentation and real-world applications where resources are limited. While transformers remain a promising direction, in this task we prioritized interpretability, efficiency, and robustness in noisy, under-resourced data conditions.

5. Experiments

5.1. Setup

Implemented in Python (scikit-learn, TensorFlow/Keras). Training used 5-fold stratified CV. Hyperparameters were tuned empirically. Experiments were run on a single TPU.

5.2. Baselines

- Majority class prediction.
- Logistic Regression + TF–IDF (text-only).

6. Results

Our models outperform baselines across all subtasks (Table 1). Vulgarity detection particularly benefited from multimodal fusion. The following table represents the evaluation metrics of the best run submission.

Table 1
HASOC 2025 Hindi memes: test set performance (macro-averaged).

Subtask	Precision	Recall	F1
Sentiment	0.71	0.69	0.70
Sarcasm	0.62	0.66	0.64
Vulgarity	0.75	0.74	0.75
Abuse	0.77	0.78	0.76

7. Discussion

Key findings:

- Preprocessing improved sentiment and abuse detection by handling noisy OCR.
- CNNs captured lexical sarcasm cues better than linear models.
- Multimodal fusion was critical for vulgarity detection, where offensiveness was primarily visual.
- Lexicon-informed features improved recall for rare abusive expressions.

8. Conclusion and Future Work

In this paper, we presented Team FAST's approach for the HASOC 2025 Hindi memes task. Our system, which placed 13th with an overall score of 0.52881, prioritized efficiency and robustness in a low-resource setting. We combined a specialized preprocessing pipeline including lexicon informed normalization for code mixing and noisy OCRwith a mix of classical machine learning (TF-IDF + Random Forest) and deep neural networks (CNN, BiLSTM, ResNet50). A key technical finding was the critical role of multimodal fusion in vulgarity detection, demonstrating that visual context is essential for classifying meme offensiveness. Our strongest individual subtask performance was recorded in abuse detection with a Macro-F1 score of 0.76. This lightweight, hybrid approach proves a valuable and practical alternative to computationally expensive transformer models for low-resource Indic language challenges, even as the overall results indicate room for improvement. Future directions could include:

- Integrating end-to-end multimodal transformers (e.g., mBERT, CLIP) for feature level fusion to potentially improve the overall standing.
- Developing and leveraging larger, more specialized Hindi, Hinglish pretrained embeddings to better handle transliteration and code-switching nuances.
- Extending the task beyond binary classification to include fine grained abuse target classification to understand *who* is being targeted in the offensive content.

Code and Resources

All code, stopword lists, offensive word dictionaries, and models are available at: https://github.com/fatimahansari/hindi-HASOC-2025

References

[1] Hasoc 2025: Hate speech and offensive content identification, https://hasocfire.github.io/hasoc/2025/, 2025. Accessed: 2025-08-18.

[2] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, A. Patel, Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages, in: Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 14–17. URL: https://doi.org/10.1145/3368567.3368584. doi:10.1145/3368567.3368584.

[3] 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.

[4] 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.

[5] 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.

[6] M. Das, S. Banerjee, A. Mukherjee, Data bootstrapping approaches to improve low resource abusive language detection for indic languages, in: Proceedings of the 33rd ACM conference on hypertext and social media, 2022, pp. 32–42.

[7] 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/.

[8] 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.

[9] 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, pp. 4171–4186. URL: https://arxiv.org/abs/1810.04805.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations (ICLR), 2021. URL: https://arxiv.org/abs/2010.11929.

[11] 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.

[12] F. A. et al., Hindi and hinglish stopword list, https://github.com/fatimahansari/hindi-HASOC-2025/blob/main/hindi_stopwords.json, 2025. Custom stopword resource used in HASOC 2025 system.

[13] F. A. et al., Hindi offensive words dictionary, https://github.com/fatimahansari/hindi-HASOC-2025/blob/main/hindi-offensive-words-original.json, 2025. Custom lexicon used in HASOC 2025 system.

[14] 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 (Eds.), Forum for Information Retrieval Evaluation (Working Notes) (FIRE 2025) December 17-20, Varanasi , India, CEUR-WS.org, 2025.