A Transformer-based approach to Multimodal Hateful Meme Classification

Sharad Prakash[1], Upkar Kumar Kedia[1], Kirti Kumari[1], Kushagra Prakash[2] and Trishant Kumar[2]

[1]Indian Institute of Information Technology, Ranchi, Jharkhand, India
[2]Amity University, Ranchi, Jharkhand, India

Abstract

In today's digital era, social media has become a common platform where people freely share their views, thoughts, ideas, and opinions. In such a scenario, it is seen that this has brought an increase in dissemination of hate speech, offensive content, derogatory remarks, etc. on such platforms. One of such hateful content is meme, which can be quickly generated and rapidly shared across online social media by users. The spread of such hateful contents has detrimental impact on people and society at large. Detecting hateful memes on such platforms is challenging due to its multimodal nature. This paper presents our approach using transformer-based model to classify the multimodal memes in four Indian languages i.e. Hindi, Gujarati, Bangla, and Bodo, for the Hate Speech and Offensive Content Identification (HASOC) 2025 shared task challenge. We participated in all the four shared subtasks and our team *KK_NLP_AI_IIIT_Ranchi*, achieved rank 3rd, 5th, 6th and 9th in Hindi, Gujarati, Bangla, and Bodo language with macro F1-score of 0.59597, 0.61409, 0.60595, and 0.57184 respectively. Further, we have discussed and compared the efficacy of the various models and architectures that we employed for the multi-task classification as presented in the challenge.

Keywords

Multimodal meme classification, HASOC 2025, Hateful meme, BERT, CLIP, IndicBERT, Feature Fusion,

1. Introduction

Due to the widespread use of social media platforms, there is an exponential growth in multimodal content, making human moderation of such information no longer feasible. In recent years, memes consist of an image with embedded text have become a very popular type of multimodal content on social networks. Memes have a significant impact on user communication styles, social culture, and content consumption patterns. Although memes are humorous in nature, they have also helped in spread of online harassment and abuse. Such hateful memes are propagated on social networks in multiple languages including English.

In South Asian countries, languages such as Hindi, Gujarati, Bangla, and other low-resource languages are prevalent. The detection of hateful memes in such languages presents unique complexities. This study aims to address these challenges by developing a model tailored to the linguistic and cultural contexts mainly of four languages i.e. Hindi, Gujarati, Bangla and Bodo by employing Natural Language Processing (NLP) techniques.

The Hate Speech and Offensive Content Identification (HASOC) shared task provides a platform for advancing research in identifying offensive and harmful content online [1]. The seventh edition, HASOC 2025, focuses on multimodal memes, which contain image with embedded text [2] to convey nuanced or implicit messages, and hence this year's challenge is organized into four main subtasks as under:

1. **Sentiment detection** – classifying memes as positive, negative, or neutral.
2. **Sarcasm detection** – identifying whether content is sarcastic or straightforward.
3. **Vulgarity detection** – distinguishing vulgar from non-vulgar memes.

Forum for Information Retrieval Evaluation, December 17-20, 2025, Varanasi, India
sharadprakash117@gmail.com (S. Prakash); upkar.2023dr101@iiitranchi.ac.in (U. K. Kedia); kirti@iiitranchi.ac.in (K. Kumari); kushagraprakash255@gmail.com (K. Prakash); trishantkumar56@gmail.com (T. Kumar)

(a) Hindi

(b) Gujarati

(c) Bangla

(d) Bodo

Figure 1: Example of memes from train datasets in four languages

4. **Abuse detection** – detecting abusive versus non-abusive content.

In this paper, we present our solution for HASOC 2025, which leverages CLIP embeddings for multimodal visual-text [3] features and BERT-based embeddings for textual content [4] [5]. We employ a simple concatenation-based fusion strategy to combine these representations for classification across all subtasks. Our architecture achieved competitive results, ranking 3rd in Hindi, 5th in Gujarati, 6th in Bangla, and 9th in Bodo language on the public test dataset shared with us. In addition, we present a comparison between multilingual and unilingual feature extraction, providing insights into their relative effectiveness for meme analysis.

An example of a meme provided with each of the four languages, i.e., Hindi, Gujarati, Bangla, and Bodo dataset, is depicted in Figure 1a, 1b, 1c, and 1d respectively. A significant number of the images used both english and the indic language for text, which further complicated the training procedure.

2. Related Works

Since meme classification is a multimodal task involving both visual and textual information, it has attracted considerable attention in recent years. Previous research has explored a variety of strategies to address the challenges of capturing the complex and context-dependent relationships between text and images, which are crucial for deciphering the intended meaning of memes.

In parallel, the HASOC (Hate Speech and Offensive Content Identification) shared tasks at FIRE (Forum for Information Retrieval and Evaluation) have established themselves as an important benchmark series for evaluating methods in hate speech detection across multiple languages. Starting with HASOC 2019 [6], which covered Indo-Euporian languages followed by HASOC 2020 with other languages such as Tamil, Malayalam, Hindi, English, and German [7], the track has expanded to include more languages and new subtracks. HASOC 2021 introduced Indo-Aryan languages and conversational hate speech [8], while HASOC 2022 further extended coverage of English and Indo-Aryan languages [9]. The 2023 edition incorporated low-resource languages such as Assamese, Bengali, Bodo, Gujarati, and Sinhala [10]. A Bengali meme dataset was created to establish an effective benchmark for classifying abusive Bangla memes using multimodal models [11]. Most recently, HASOC 2024 focused on hate speech detection in English and Bengali [12]. Ghosh et al. [13] explored a novel three-module pipeline, SafeSpeech, for Hate Speech Classification, Hate Intensity Identification, and Hate Intensity Mitigation using publicly available datasets in five Indic languages (Hindi, Marathi, Tamil, Telugu, and Bengali). Recently, Ghosh et al. [14] developed models on machine learning and deep learning methods for the detection of Hate speech in four low-resource languages (Assamese, Bengali, Bodo, and English) datasets, and analyzed their performance across standard evaluation metrics parameters. Their findings offered challenges and progress in the detection of hate speech in low-resource multilingual languages. Furthermore, Vijay et at. [15] develop a hate speech detection framework by employing TF-IDF word embedding technique for feature extraction and a lexicon-based hierarchical approach to address challenges of linguistic and cultural intricacies in Hinglish (a mix of Hindi and English) and Bangla languages.

Table 1

Details of HASOC 2025 dataset

Language	Training	Testing	Total
Hindi	1141	769	1910
Gujarati	889	604	1493
Bangla	2693	1821	4514
Bodo	378	254	632

Our work builds on this line of research in HASOC Track at FIRE 2025 on abusive Meme identification [16], [17], extending the scope to multimodal meme classification and advancing hate speech detection in multilingual and multimodal settings.

3. Dataset Description

The HASOC 2025 shared task released training and test datasets for four languages: Hindi, Gujarati, Bangla, and Bodo (see Table 1). The distribution of training datasets across the four languages was highly unbalanced, ranging from just 378 memes for Bodo to 2693 dataset points for Bangla. Due to this disparity in the size of the data sets, we adopted language-specific training strategies to better account for the varying levels of data availability.

The datasets included an additional column named 'Target', which was eventually omitted from the sample submission and hence removed from our training data. The Bodo training dataset also had a typographical error in the 'Vulgar' column, which was coerced to 'Non-Vulgar' class and corrected.

It was also observed that for Bodo, the images generally did not provide much semantic context to the actual meaning of the meme, and even worsened the results when used with the Optical Character Recognition (OCR) text for training. In contrast to this, the Bangla images improved significantly when the images were coupled with the OCR text during training.

There was no separate validation dataset provided, hence we used a 90:10 Train:Validation split on our train set.

4. Methodology

This section describes an overview of feature extraction techniques for text-only model and Contrastive Language-Image Pre-Training (CLIP) [18] based model with text and the methods used to train these models.

4.1. Text-only Models

This section deals with our initial attempts at the classification problem with text-only, low-resource models that used only the OCR text for feature extraction.

4.1.1. Multilingual Backbone (IndicBERT)

We first experimented with a multilingual backbone by fine-tuning the **IndicBERT** model, trained on a large corpus of Indian languages by AI4Bharat. This purely text-based architecture, as in Figure 2, surpassed the baseline by a large margin.

For the Bodo, Hindi, and Gujarati datasets, this setup achieved macro F1-scores greater than 0.55. However, the Bangla dataset, which had more data points than the other three languages combined, performed poorly, with the macro F1-score dropping well below the baseline of 0.53. We attribute this to the presence of a large number of images in the Bangla dataset that provided crucial semantic context to the memes, which was not captured by a text-only architecture.

This multilingual, text only model architecture achieved a test macro F1-score of 0.56541 for Hindi, 0.56775 for Gujarati, 0.54691 for Bodo and 0.45601 for Bangla.

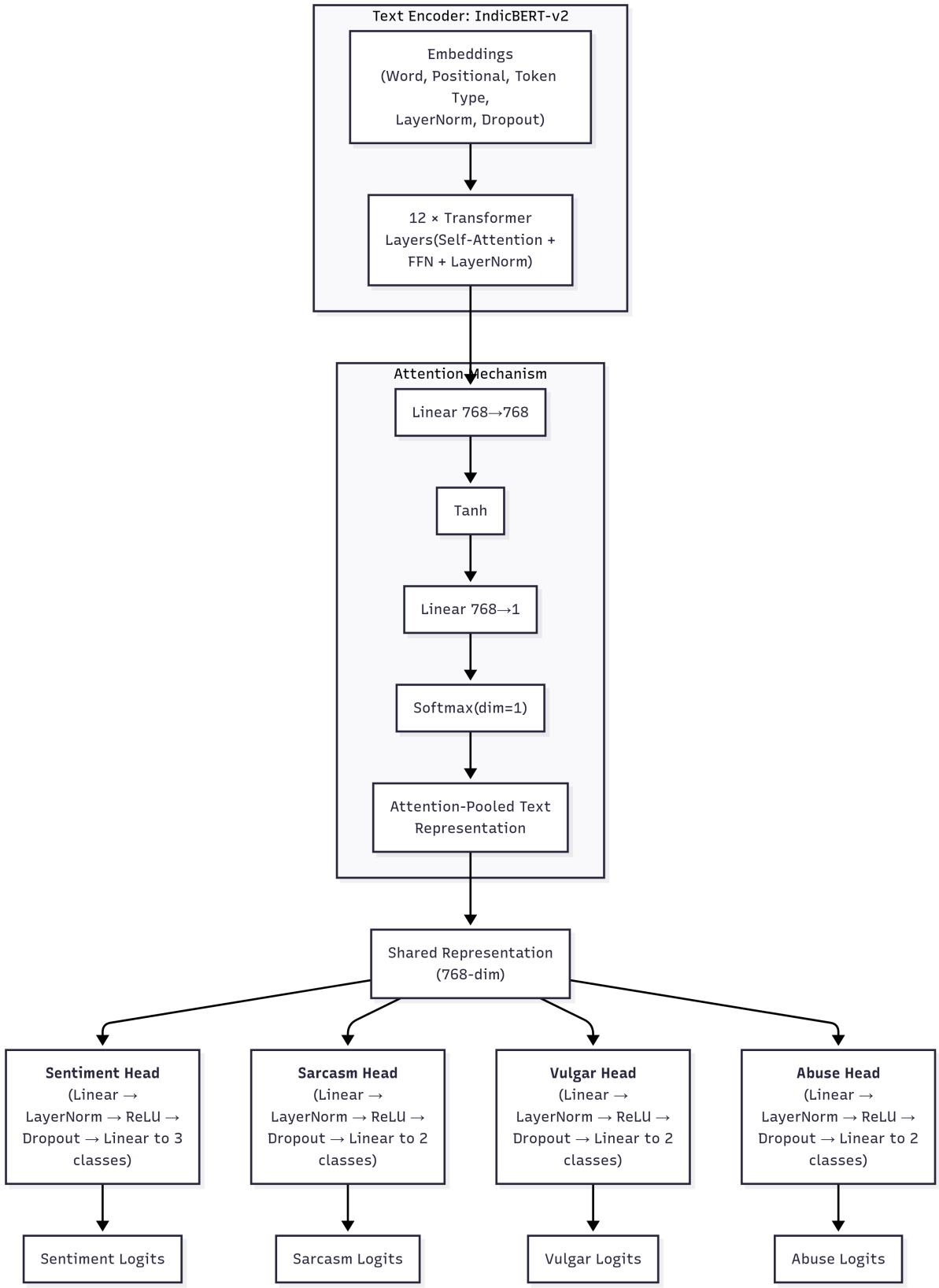


Figure 2: Text only model architecture

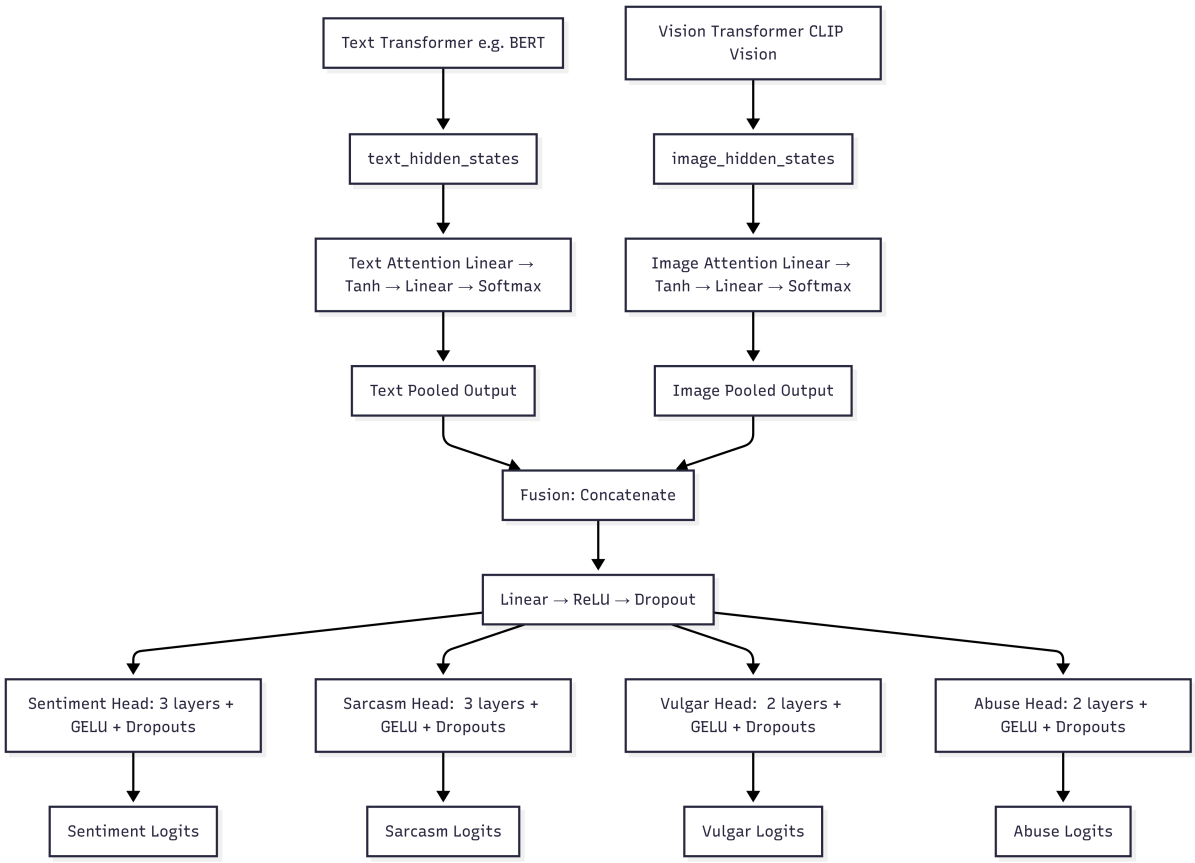


Figure 3: CLIP with Text feature fusion model

4.1.2. Monolingual Backbones

In order to improve the performance of the models and to ensure that the poor results on Bangla were not due to limitations of the IndicBERT backbone, we also experimented with monolingual BERT-based models specific to each language [19]. We used **bengali-bert** for Bangla, **hindi-bert-v2** for Hindi, **gujarati-bert** for Gujarati, and **pretrained-bodo-legal-bert** for Bodo.

These monolingual models did not result in any significant improvements for the Bangla, Hindi or Gujarati models, each of them staying close the previous results achieved by IndicBERT. The bodo model however, worsened even further, potentially due to poor training corpus of the backbone model used.

4.2. CLIP with Text models

In order to ensure the usage of the image features with the text features extracted using the BERT based backbone, we incorporated the CLIP Vision model [3] as shown in Figure 3. For simplicity, we performed a simple feature concatenation to fuse the two features; however, later investigation using cross-attention-based feature concatenation only worsened the results.

The CLIP model being trained primarily on the English language failed to correlate and extract join image-text features for the Indic languages, so cross-attention-based feature fusion was omitted from the architecture and only simple concatenation-based feature fusion was used in all the subsequent models.

As was the case with the text-only models, we experimented with both multilingual and monolingual backbones for text feature extraction. For Bangla, the multilingual, IndicBERT-based backbones coupled

with CLIP vision model, improved the results significantly, improving the test macro F1-score to 0.58 from 0.45. The Gujarati model too, improved its macro F1-score to 0.59 from 0.56 with this multimodal model. However, Hindi did not benefit from IndicBert and CLIP, with results staying close to the previous text only models, and Bodo worsened on using the CLIP features. This poor performace for Bodo remained consistent across both monolingual and multilingual text backbones, hence the image features were dropped for Bodo during final submission. One of the potential reasons for this was the very low number of data points for Bodo (378), where the majority of the images did not provide any semantic context to the meme. The final submission for Bodo was made using the IndicBert-v2 model trained using the text only architecture, with a Test macro F1-score of 0.57184.

The monolingual backbones did improve the results on Hindi, Gujarati and Bangla, surpassing the macro F1 threshold of 0.60 for both Bangla and Gujarati, and achieving a macro F1-score of 0.59597 for Hindi. The final scores for the submission were achieved using the CLIP and monolingual models for Bangla, Gujarati and Hindi, with test macro F1-scores of 0.60595, 0.61409, and 0.59597 respectively.

4.3. Training

All models were trained on Kaggle servers using a P100 GPU. During training, we observed that the training curves behaved quite aberrantly, even with the Adam optimizer. To mitigate this and ensure stable convergence, we employed several techniques. We utilized weight decay and a warm-up based learning rate scheduler. Furthermore, early stopping was implemented to prevent overfitting and select the best-performing models for final prediction.

Two distinct training strategies were employed based on language-specific performance:

1. **Cumulative Training:** In this method, all classification heads were trained simultaneously, sharing a common feature pool from the backbone. This approach proved most effective for the Bodo model, which benefited from the shared feature representation. When trained separately, the Bodo model's loss and macro F1 scores oscillated and failed to converge.
2. **Sequential Training:** Here, each classification head was trained separately while the other layers were frozen. This sequential process yielded better results for Hindi, Gujarati, and Bangla.

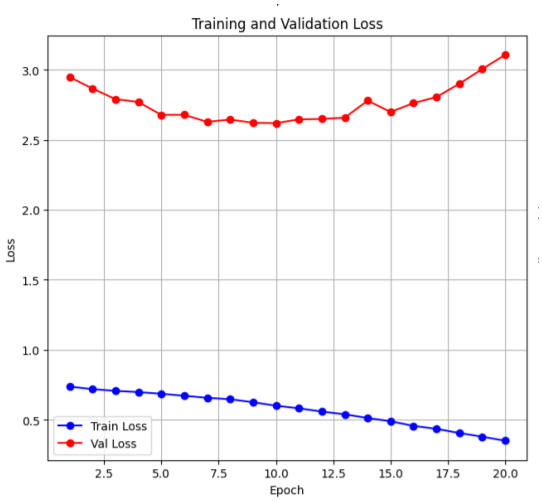
5. Results and Discussion

This section presents the results and discussion of our multimodal hateful meme classification approach, analyzing performance of our model across the four shared languages, i.e. Hindi, Gujarati, Bangla, and Bodo. The macro F1 scores of our model for the four languages compared with the first-ranked models and our ranking in the participation are mentioned in Table 2. Plots for Loss and macro F1-score are illustrated in Figure 4 and Figure 5 respectively.

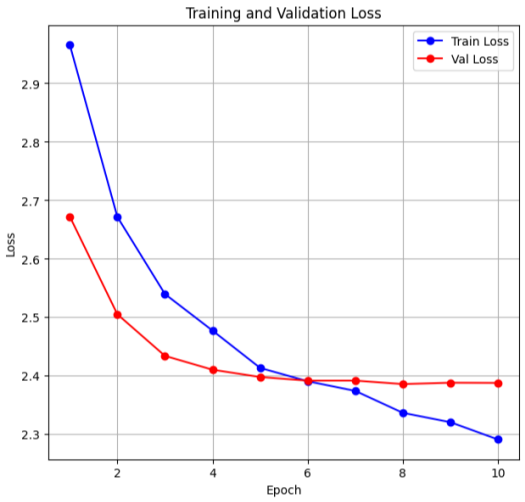
Table 2
Macro F1-scores of our model compared with the first-ranked model and our rank in the HASOC 2025 challenge task

Language	Model	Macro-F1	Rank 1	Our Rank	Number of submissions
Hindi	CLIP with BERT	0.59597	0.65706	3	18
Gujarati	CLIP with BERT	0.61409	0.67501	5	15
Bangla	CLIP with BERT	0.60595	0.62755	6	17
Bodo	IndicBERT	0.57184	0.63128	9	15

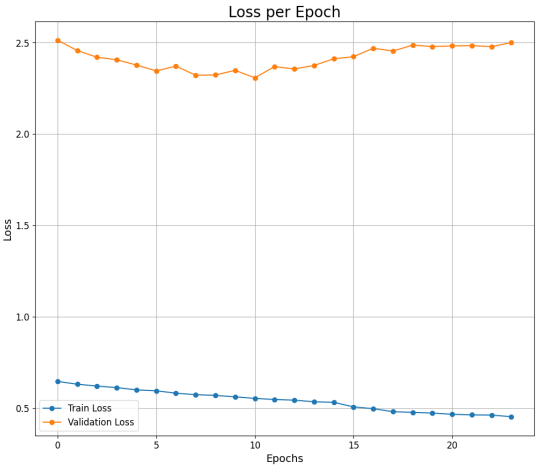
The Gujarati model converged quite well, as can be seen from the macro F1-score and loss plots (Figure 4 and 5). However, Bodo, Bangla, and Hindi models required the careful scheduling of learning rates to achieve convergence. Despite these efforts, the Bodo model did not converge properly; even with learning rates as low as 5e-6, the loss values continued to oscillate during training. Hindi, Bangla and Bodo, all overfit quite quickly, in contrast to Gujarati, which achieved the highest test macro F1 score across all the languages.



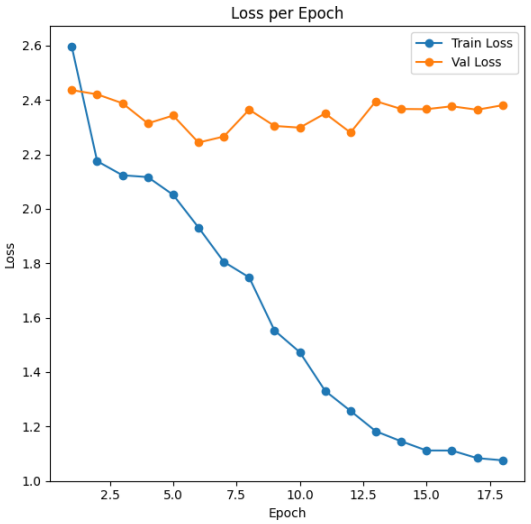
(a) Hindi



(b) Gujarati



(c) Bangla



(d) Bodo

Figure 4: Train and validation loss for all four languages

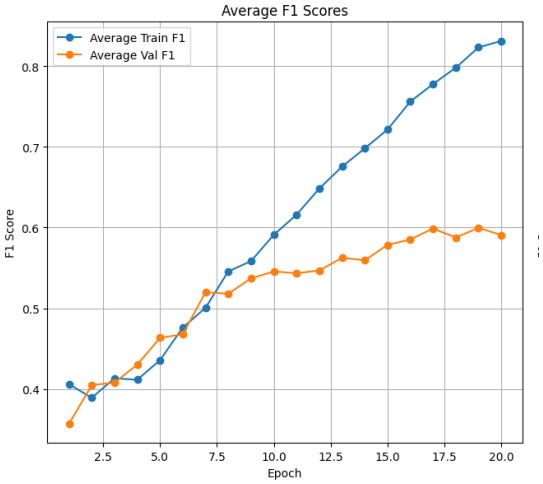
When comparing the training and validation macro F1-scores as seen in the Figure 5, the Bodo model had a huge difference between the training and validation macro F1-scores compared to the rest of the languages. The Bodo model quickly overfit on the small corpus of training set, and hence the poor result on the validation set.

It was also observed that the "Vulgar" and "Abuse" classification heads converged the fastest, generally achieving validation macro F1-scores above 0.70 across all languages. This correlation was equally present with the "Sentiment" and "Sarcasm" heads which converged slowly and had validation macro F1-scores that stayed close to each other.

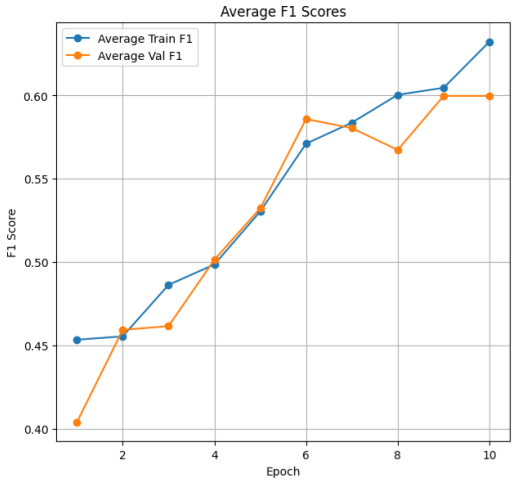
Our models ranked 3rd out of 18 submissions for Hindi, 5th out of 15 submissions for Gujarati, 6th out of 17 submissions for Bangla, and 9th out of 15 submissions for Bodo.

6. Conclusion and Future Work

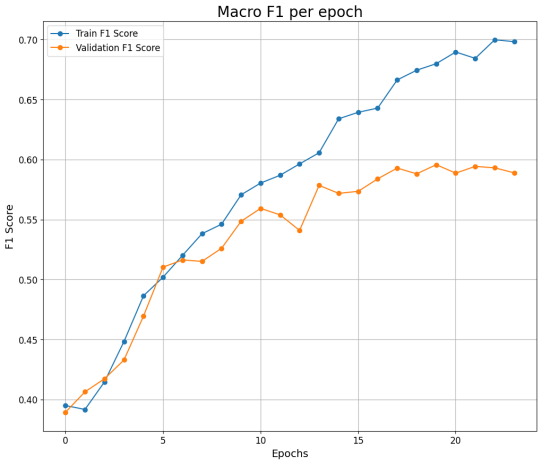
In this work, we presented a comprehensive approach to multi-label offensive meme classification for four Indic languages. Our findings underscore the critical importance of multimodality, where fusing



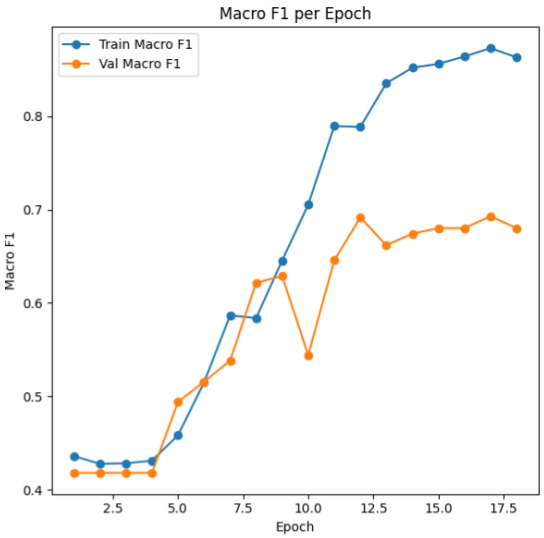
(a) Hindi



(b) Gujarati



(c) Bangla



(d) Bodo

Figure 5: Macro F1 scores across all four languages

visual features from CLIP with monolingual BERT backbones yielded significant performance gains for Hindi, Gujarati, and Bangla. Conversely, for the low-resource Bodo language, a text-only model proved superior, highlighting that the optimal architecture is highly dependent on the dataset's characteristics. Our models achieved competitive ranks in the HASOC 2025 challenge, validating our methodology.

Based on the challenges that we observed regarding multimodal feature fusion and low resource languages, we propose moving beyond simple feature concatenation by integrating inherently multilingual vision-language models like **mclip** [20], which could provide better-aligned representations and enable more sophisticated fusion mechanisms. This would be especially crucial for improving performance on low-resource languages like Bodo, where techniques such as few-shot learning and advanced data augmentation could also be explored. Furthermore, to capture a richer semantic context from memes, future models could incorporate visual text features beyond simple OCR, leveraging layout-aware architectures like **LayoutLM** [21] to understand the stylistic nuances of text within the image.

7. Acknowledgments

As participants in the HASOC 2025 Challenge, we fully comply with the competition rules as outlined by the organizers on the challenge website. Our methods have used the training and test data sets provided in the official release of the datasets to report the results of the challenge.

Declaration on Generative AI

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

References

[1] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P.-E. Mazaré, D. Ju, J. Chang, W. Galuba, C. Musat, et al., The hateful memes challenge: Detecting hate speech in multimodal memes, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 2626–2639.

[2] E. Hossain, O. Sharif, M. M. Hoque, S. M. Preum, Deciphering hate: Identifying hateful memes and their targets, 2024. URL: https://arxiv.org/abs/2403.10829. arXiv:2403.10829.

[3] S. B. Shah, S. Shiwakoti, M. Chaudhary, H. Wang, Memeclip: Leveraging clip representations for multimodal meme classification, 2024. URL: https://arxiv.org/abs/2409.14703. arXiv:2409.14703.

[4] M. V. Koroteev, Bert: a review of applications in natural language processing and understanding, arXiv preprint arXiv:2103.11943 (2021).

[5] J. Devlin, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[6] 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.

[7] T. Mandl, S. Modha, A. Kumar M, B. R. Chakravarthi, Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german, in: Proceedings of the 12th annual meeting of the forum for information retrieval evaluation, 2020, pp. 29–32.

[8] 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, 2021, pp. 1–3.

[9] S. Satapara, P. Majumder, T. Mandl, S. Modha, H. Madhu, T. Ranasinghe, M. Zampieri, K. North, D. Premasiri, Overview of the hasoc subtrack at fire 2022: Hate speech and offensive content identification in english and indo-aryan languages, in: Proceedings of the 14th annual meeting of the forum for information retrieval evaluation, 2022, pp. 4–7.

[10] 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.

[11] 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.

[12] T. Mandl, K. Ghosh, N. Raihan, S. Modha, S. Satapara, T. Gaur, Y. Dave, M. Zampieri, S. Jaki, Overview of the hasoc track 2024: Hate-speech identification in english and bengali, in: Proceedings of the 16th Annual Meeting of the Forum for Information Retrieval Evaluation, 2024, pp. 1–2.

[13] 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.

[14] 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 199 (2025) 303–309. URL: https://www.sciencedirect.com/science/article/pii/S0167865525003150. doi:https://doi.org/10.1016/j.patrec.2025.09.004.

[15] V. Vijay, A. Bhattacharjee, K. Kumari, U. K. Kedia, Detecting hate speech in bangla: A hybrid model using machine learning and lexicon-based strategies (2024).

[16] 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), CEUR-WS.org, Varanasi, India, 2025.

[17] 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, Association for Computing Machinery (ACM), Varanasi, India, 2025.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[19] D. Kakwani, A. Kunchukuttan, S. Golla, C. Gokul, A. N, S. Jain, S. Jain, R. Patel, T. V. S. L. S. Kumar, D. Shah, et al., IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 4461–4473.

[20] G. Chen, L. Hou, Y. Chen, W. Dai, L. Shang, X. Jiang, Q. Liu, J. Pan, W. Wang, mclip: Multilingual clip via cross-lingual transfer, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 13028–13043.

[21] Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, Y. Lu, D. Florencio, C. Zhang, W. Che, M. Zhang, L. Zhou, Layoutlmv2: Multi-modal pre-training for visually-rich document understanding, 2022. URL: https://arxiv.org/abs/2012.14740. arXiv:2012.14740.