FusionGuard: Visual-Linguistic Representations for Multilingual Harmful Content Detection

Arunabha Basak[1,†], Samujjal Choudhury[1,†], Anshita Malviya[1,*,†], Bidyut Kr. Patra[1] and Pratik Chattopadhyay[1]

[1]*Indian Institute of Technology (BHU) Varanasi, India*

Abstract

The growth of internet technologies has given rise to various social media platforms such as Twitter, Facebook, Instagram, and many more. People all over the world, irrespective of their age, have access to these platforms and can use them to upload their day to day events. Everyone has the power to voice their opinions on various items/posts of other people as well. Although this helps to bring people together, sometimes this can lead to issues within communities and groups. The vast majority of people tend to post offensive/hate content on the internet. This has become a serious global concern. To solve this problem, a method needs to be introduced which can analyze huge amounts of data (text and images) over the internet and filter out the offensive content. Our paper introduces 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 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. Within the HASOC 2025-Meme task, our Team DeepSemantics ranked 11th, 14th, 11th, and 17th in Bodo, Gujarati, Hindi and Bangla datasets respectively. We achieved F1-scores of 0.5604 on Bodo, 0.4203 on Gujarati, 0.5498 on Hindi, and 0.4821 on Bangla datasets.

Keywords

Multimodal classification, ResNet, VisualBERT, Image embeddings, Text embeddings

1. Introduction

Social media platforms have become an integral part of modern communication, enabling users to share content and express their opinions on a global scale. While these platforms were originally intended to foster open dialogue and the exchange of ideas, they have also become channels for the spread of negativity, hate speech and offensive content [1]. The prevalence of such harmful communication poses significant challenges for platform providers, who are under increasing pressure to detect and mitigate abusive behavior effectively.

Hate speech refers to derogatory expressions or terms directed at individuals or groups with the intention to cause harm. Such expressions are typically based on characteristics such as ethnicity, gender, religion, disability, sexual orientation, or nationality. Hate speech can have harmful consequences at both individual and societal levels, fostering discrimination, hostility, and social division, and therefore requires effective mechanisms for detection and removal [2] .

Nowadays, memes have become a frequent mode of online communication, combining text and images to convey humorous or satirical messages. However, some memes may also contain hateful or offensive content. The multimodal nature of memes makes them difficult to analyze, which often causes traditional text-based hate speech detection models to fail. Detecting harmful content in memes

Forum for Information Retrieval Evaluation, December 17-20, 2025, Varanasi, India

*Corresponding author.

†These authors contributed equally.

arunabhabasak.rs.cse23@iitbhu.ac.in (A. Basak); samujjalchoudhury.rs.cse24@iitbhu.ac.in (S. Choudhury); anshitamalviya.rs.cse23@iitbhu.ac.in (A. Malviya); bidyut.cse@iitbhu.ac.in (B. Kr. Patra); pratik.cse@iitbhu.ac.in (P. Chattopadhyay)

0009-0003-6483-2674 (A. Basak); 0009-0004-7084-2104 (S. Choudhury); 0009-0004-8647-0245 (A. Malviya); 0000-0003-3012-4285 (B. Kr. Patra); 0000-0002-5805-6563 (P. Chattopadhyay)

therefore requires approaches that consider both textual and visual modalities, posing a significant global challenge.

The Hate Speech and Offensive Content Identification (HASOC) is a shared task series which is organized anually as part of the Forum for Information Retrieval Evaluation (FIRE) since 2019 [3]. The main goal of HASOC is to find the best approaches for finding hate speech, offensive language and abusive content from social media data. HASOC emphasizes multilinguality and has progressively included datasets in various languages other than English. The HASOC benchmark has provided a consistent and challenging platform for advancing robust, multilingual and multimodal hate speech detection systems, with wide adoption in the natural language processing and social media analysis communities.

The 2025 edition of HASOC significantly extends previous years' efforts by introducing multi-task, multilingual, and multi-dimensional classification challenges, reflecting the complex nature of harmful communication on social media platforms. Specifically, HASOC 2025 addresses the classification of social media posts (memes) across four different languages and for four distinct application domains:

- Sentiment analysis: Identifying the polarity of online expressions to capture effective attitudes conveyed in user-generated content.
- Sarcasm Detection: Detecting sarcastic expressions, which pose unique challenges for computational models due to their reliance on contextual cues, implicit meaning and irony.
- Abuse Detection: Classifying abusive and derogatory language that targets individuals or groups, thereby aiming to protect users from direct harm and online harassment.
- Vulgarity Identification: Identifying the use of vulgar, profane, or explicit language, which-although not always hateful contributes to offensive and harmful communication.

HASOC encourages the development of generalizable and robust NLP models capable of handling diverse manifestations of offensive content. By framing the task across multiple languages and content categories, HASOC 2025 provides a comprehensive benchmark to evaluate multilingual, cross-domain and multimodal approaches in offensive content detection.

The datasets provided for HASOC 2025 are sourced from real-world social media platforms, ensuring both practical relevance and linguistic diversity. This shared task offers researchers an opportunity to develop models that are not only accurate but also sensitive to cultural and linguistic nuances.

The remaining sections of the paper are structured as follows. In Section 2, a survey of techniques for classification tasks is provided. The dataset description is presented in Section 3. The proposed architecture has been discussed in Section 4. Results and discussions are presented in Section 5, and the conclusion is provided in Section 6

2. Related Work

Due to the rise of multimedia content on the internet, multimodal (text and images) analysis has emerged as an important and critical area of research. The various classification tasks mentioned in previous section not only need textual data but also images for better understanding of contexts. For example, a comment may convey a positive sentiment but its image may represent a sarcastic tone. Also, abusive and vulgar content may rely on visual representations which might not be detected from text alone.

Hate speech and abusive content detection in Indic languages has been an active research area, with multiple shared tasks organized under HASOC in recent years to support benchmark development and multilingual evaluation [4][5][6]. Further work has introduced datasets and model improvements for low-resource languages such as Bengali and other Indic scripts [7][8]. The HASOC 2025 Abusive Meme Identification track extends this focus to multimodal settings involving memes, combining both visual and textual cues [9][10]. Our work builds upon these initiatives by evaluating multimodal architectures for multilingual meme classification across four Indic languages.

Chanda et al. (2021) fine-tuned pre-trained transformer models across English, Hindi, Marathi and English-Hindi code-mixed data within the HASOC 2021 shared task framework, which illustrated their

versatility in cross-lingual and mixed-language scenarios [11]. At SemEval-2020, the IRlab@IIT-BHU team applied an SVM classifier with TF-IDF features to detect offensive language in multiple languages [12]. The "Crossing Borders: Multilingual Hate Speech Detection" paper examines sentence-level hate/offensive detection in less-resourced languages such as Gujarati and Sinhala, emphasizing the challenges inherent in adapting models across diverse linguistic landscapes [13]. In HASOC 2022, IRLab@IITBHU submitted a model based on fine-tuning of existing models like XLM Roberta and German BERT for classification of tweets in Marathi, Hinglish Codemix and German language [14] . A deep learning approach to Hindi-English code-mixed hate speech detection explores sequential modeling architectures tailored to capture the nuances of mixed-language text in social media contexts [15]. In addition, Hate Content Identification in Code-Mixed Social Media Data by Chanda and Pal (2022) investigates Hindi-English code-mixed conversations, showing that deep learning architectures can effectively capture linguistic variation and identify hate content in noisy real-world contexts [16]. Collectively, these studies demonstrate that transformer-based models, particularly when adapted for multilingual and code-mixed data, offer notable advantages in accuracy and adaptability over traditional machine learning classifiers.

Two major approaches exist for multimodal classification tasks. One is by making use of convolutional neural networks (CNNs) [17] and the other is through transformer based models [18]. CNNs are capable of extracting high quality visual features and transformer architectures are good for joint reasoning on image-text pairs.

Traditional CNN models such as AlexNet [19], VGGNet [20] and GoogLeNet [21] had a common issue known as the vanishing/exploding gradient problem. Also, it was observed that the accuracy of such models would remain saturated or even drop if more layers were introduced. To solve these issues, the ResNet (Residual Network) architecture [22] was introduced which implemented skip connections. These would prove effective in training deep networks consisting of multiple layers. Instead of directly learning a mapping $H(x)$ from input x to output, ResNet learns a residual mapping $F(x) = H(x) - x$. The final output is computed as $y = F(x) + x$, where x is propagated through the network via an identity shortcut connection. Common ResNet variants include ResNet-18, ResNet-34 and ResNet-50. These differ primarily in depth and the type of residual blocks used.

A joint understanding of images and text is necessary in multimodal domain. A powerful transformer model called BERT (Bidirectional Encoder Representations from Transformers) [23] revolutionized the domain of Natural Language Processing. However, it was designed to handle only text data. VisualBERT [24] is a model that injects image features into the BERT's transformer architecture along with the textual features. This enables reasoning over both text and image modalities.

3. Datasets

The HASOC 2025 shared task comprises datasets in four different languages: Hindi, Bengali, Bodo and Gujarati. For each language, the dataset is designed to support four distinct classification tasks: Sentiment Analysis, Sarcasm Detection, Vulgarity Detection and Abusive Content Classification. This multi-task, multilingual setup provides a comprehensive benchmark for evaluating natural language processing models in low-resource and high-resource language contexts. The detailed statistics of each dataset is presented in Table 1. For the Sentiment and Vulgarity classification tasks, the labels are represented as -1 (Negative), 0 (Neutral), and 1 (Positive). In contrast, for the Sarcasm and Abuse detection tasks, a binary labeling scheme is adopted, where 0 denotes Non-Sarcastic/Non-Abusive and 1 denotes Sarcastic/Abusive.

4. Proposed Methodology

As described in previous section, the datasets consist of images as well as the extracted text. In this paper we utilize ResNet and VisualBERT to perform multimodal classification across all languages.

Table 1
Statistics of the HASOC 2025 datasets across four languages for Sentiment Analysis, Sarcasm Detection, Vulgarity Detection and Abusive Content Classification.

Language	Sentiment (-1 / 0 / 1)			Sarcasm (0 / 1)		Vulgar (-1 / 0 / 1)			Abusive (0 / 1)	
	-1	0	1	0	1	-1	0	1	0	1
Hindi	525	276	340	371	770	764	0	377	834	307
Bengali	1476	311	906	612	2081	2226	0	467	1954	739
Bodo	151	0	227	39	339	269	2	107	301	77
Gujarati	291	194	404	219	670	592	0	297	721	168

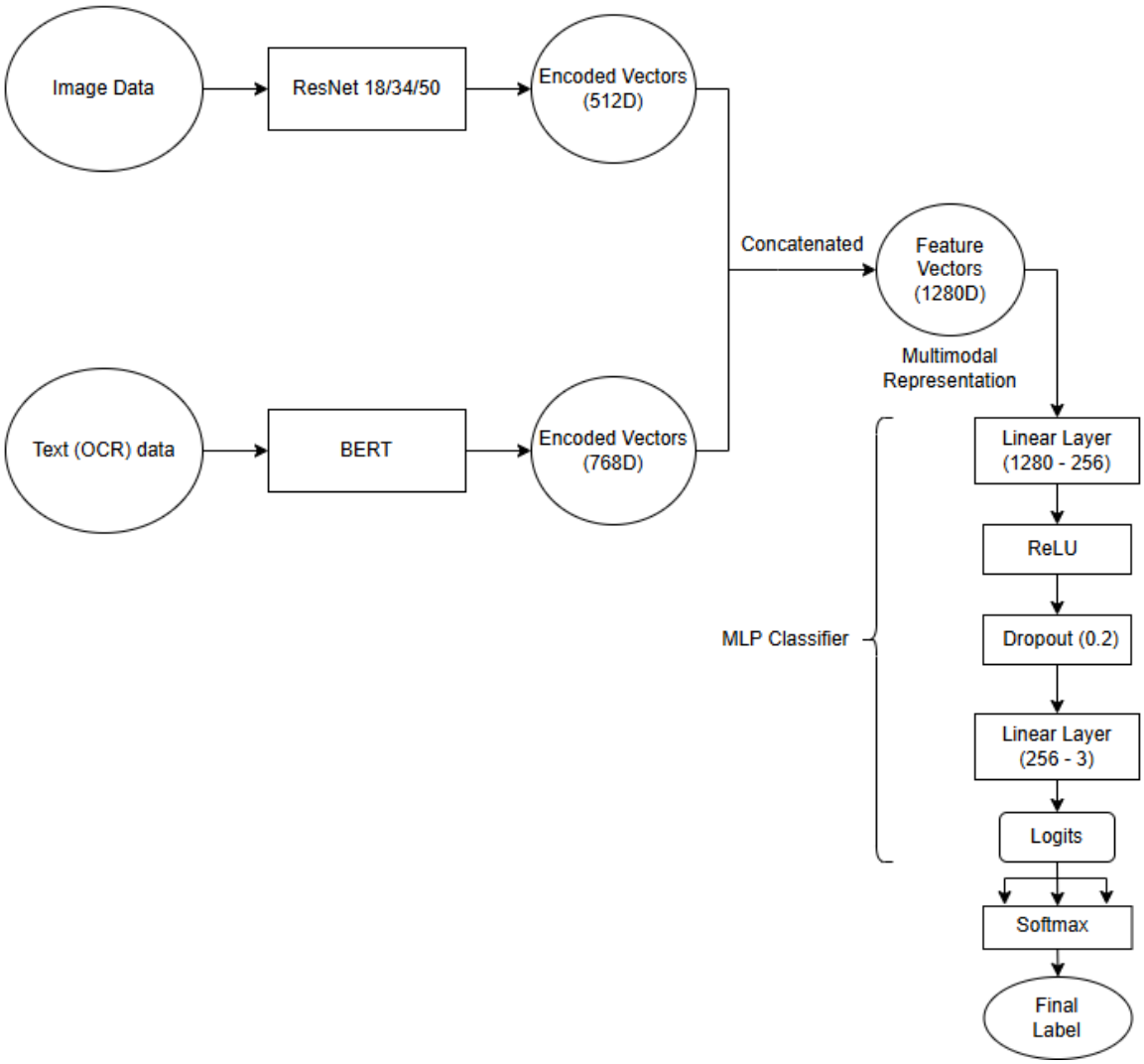

Figure 1: Architecture of the proposed multimodal classifier with ResNet for three output labels.

4.1. ResNet multimodal architecture

In the multimodal architecture using ResNet, a pre-trained ResNet backbone (ResNet-18/34/50) is used for visual modality. This network processes images and outputs vector embeddings in the form of 512-dimensional feature vectors. These embeddings capture both low-level patterns and high-level semantic features from the image domain.

To extract the features of textual data, the OCR extracted texts are encoded using BERT. It converts the textual input into 768-dimensional embedding vectors. The resulting vector effectively represents the

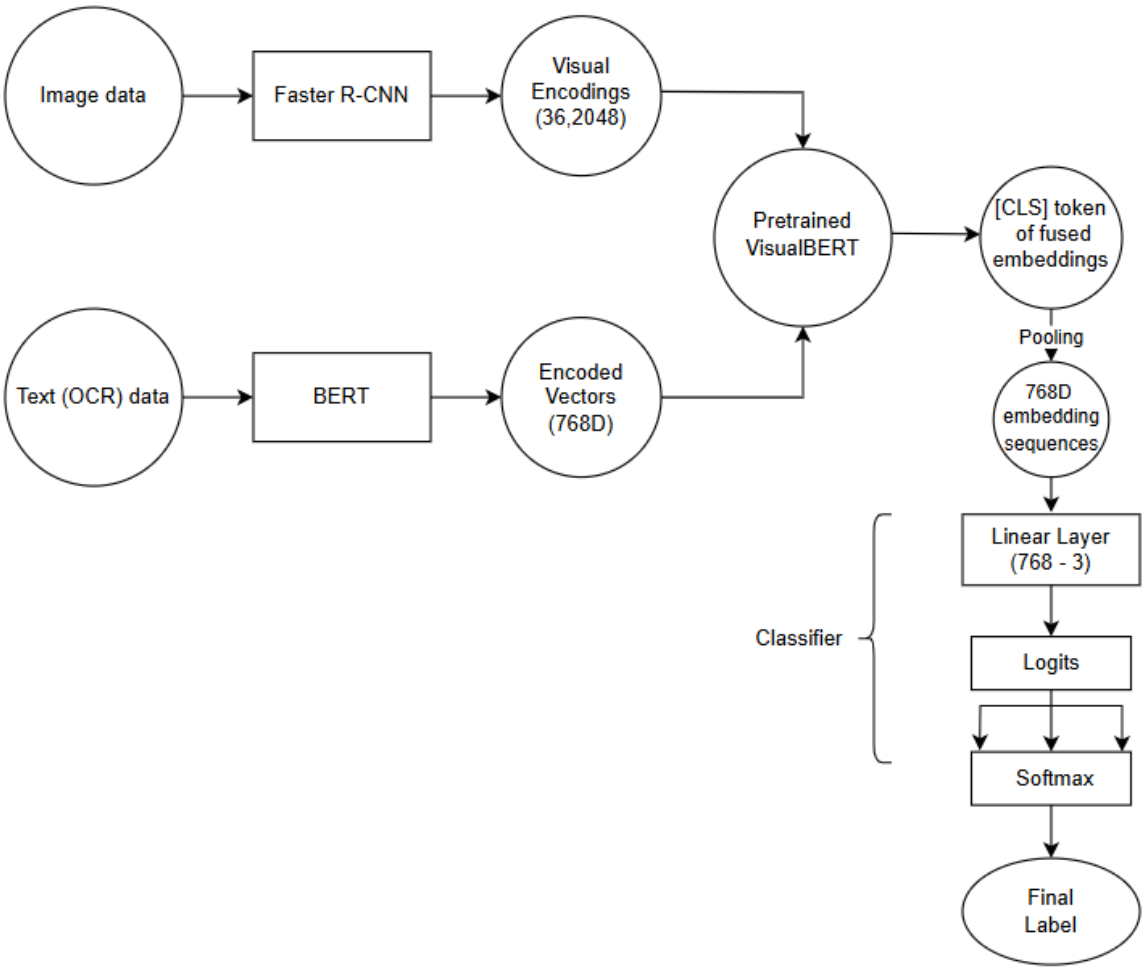

Figure 2: Architecture of the proposed multimodal classifier with VisualBERT for three output labels.

semantic information contained in the text, preserving both word-level and contextual dependencies.

The visual and textual embeddings are concatenated to form a unified 1280-dimensional multimodal embedding vectors. This representation is then passed into a classifier (Multilayer Perceptron). The classifier consists of a linear transformation, followed by a ReLU activation and dropout to prevent overfitting. The output is then passed through a final linear layer to produce logits corresponding to target classes. A softmax function is applied to the logits to generate the final class probabilities.

The entire model is trained end-to-end, enabling the joint optimization of both modalities and ensuring that the fused representation captures complementary visual and textual cues for robust classification. The proposed architecture using ResNet is shown in figure 1.

4.2. VisualBERT multimodal architecture

In the multimodal architecture utilizing VisualBERT, Faster R-CNN (Regions with Convolutional Neural Networks) is used as an object detector to capture semantically meaningful regions of interest (ROIs). This model generates a set of visual embeddings for an input image and each embedding is represented as a 2048-dimensional vector. The top 36 detected regions are used for computational efficiency and the output becomes a matrix of shape (36,2048).

For the textual embeddings, the extracted text from the images are encoded using a pretrained BERT model. This model maps tokenized sequences into 768-dimensional embedding vectors.

After obtaining the image and textual embeddings, a pretrained VisualBERT is utilized to align the 36 visual embeddings with the text embeddings in a 768-dimensional feature space. This is achieved

Table 2
Accuracy comparison of ResNet and VisualBERT architectures on multilingual meme classification

Language	Model	Sentiment	Sarcasm	Abuse	Vulgar
Gujarati	ResNet-18	**0.4382**	0.6404	0.8034	**0.7079**
	ResNet-34	0.4213	0.6685	0.7980	0.6348
	ResNet-50	0.4157	0.6798	0.76278	0.5674
	VisualBERT	0.4338	**0.7191**	**0.8146**	0.6798
Bengali	ResNet-18	0.5603	0.7199	**0.7180**	0.7848
	ResNet-34	0.5455	0.6846	0.6957	0.7904
	ResNet-50	0.5603	0.6401	0.7032	0.7625
	VisualBERT	**0.5720**	**0.7673**	0.6549	**0.7904**
Bodo	ResNet-18	**0.5789**	0.8684	0.6842	0.5000
	ResNet-34	0.5789	0.8684	0.6711	0.4605
	ResNet-50	0.5658	0.7895	0.6974	0.5921
	VisualBERT	0.5263	**0.8684**	**0.7237**	**0.6974**
Hindi	ResNet-18	0.4672	0.6463	0.6856	**0.6987**
	ResNet-34	0.4323	0.5328	0.7074	0.6157
	ResNet-50	**0.4847**	0.5677	0.6812	0.6419
	VisualBERT	0.4672	**0.7467**	**0.7511**	0.6725

Table 3
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 4
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

by linearly projecting the visual features to match the BERT embedding size before fusion. Both modalities are then concatenated and processed by the transformer layers of VisualBERT, which learns the cross-modal interactions. The [CLS] token of the fused sequence is used as a compact multimodal representation.

The final part of the architecture constitutes a classifier. The pooled [CLS] representation is passed through a fully connected linear layer, producing logits over three target classes. A softmax layer is then applied to obtain the final label distribution. The entire model is fine-tuned end-to-end, enabling both the visual and textual encoders to adapt to the downstream task. The entire architecture is presented in Figure 2.

Table 5
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 6
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. Results

Since only one model can be deployed for HASOC 2025, a comparative study has been conducted between ResNet (18,34,50) and VisualBERT. The initial datasets released by the HASOC team for four languages (containing labels) was divided into training and test datasets for evaluation in the ratio 80:20. To maintain fairness, all models were trained with the same hyperparameters: 10 epochs and a learning rate of 2e-5. The results obtained for all tasks have been shown in table 2.

- Gujarati: VisualBERT outperforms all ResNet models in sarcasm and abuse detection. Although ResNet-18 achieves slightly higher accuracy in sentiment and vulgarity detection, VisualBERT maintains competitive performance while excelling in more context-dependent tasks.
- Bengali: VisualBERT achieves the best sentiment and sarcasm detection, outperforming all ResNet variants. For vulgarity detection, VisualBERT ties with the best ResNet model. Abuse detection shows mixed results, but the gap is small.
- Bodo: VisualBERT achieves the highest accuracy in abuse and vulgarity detection. Although ResNet-18 and ResNet-34 outperform VisualBERT in sentiment, they all give the same result for sarcasm.
- Hindi: While ResNet models achieve slightly higher results in vulgarity and sentiment classification, VisualBERT clearly dominates in sarcasm and abuse detection.

Across all languages, VisualBERT consistently demonstrates superior performance compared to the ResNet variants (10 out of 16 tasks) , particularly in tasks that require deeper semantic understanding such as sarcasm and abuse detection. Its performance advantage over ResNet highlights the importance of transformer-based multimodal architectures, which can effectively capture nuanced relationships between textual and visual features in memes.

The variation in performance across languages for the same task is expected and can be attributed to multiple dataset and model-related factors. The datasets we have evaluated on are highly imbalanced, with Bodo and Gujarati being very small in size. This directly impacts model generalization due to lack of labeled data. Secondly, meme structure and language of the datasets vary significantly with different levels of code-mixing and script style. VisualBERT relies heavily on informative textual signals, (which are stronger in Bengali and Hindi) leading to higher scores. Also, all models were trained with the same hyperparameters to maintain fairness, without language specific tuning. While this ensures proper

comparability, it may not fully optimize performance for every language dataset, resulting in imbalance of the scores.

Our run submission consists of the results obtained from the VisualBERT architecture. Table 3, Table 4, Table 5 and Table 6 shows the overall leaderboard results of our model compared to other performing teams of HASOC 2025 in Gujarati, Bodo, Bangla and Hindi datasets respectively.

6. Conclusion

In this paper, we investigated multimodal classification approaches for detecting sentiment, sarcasm, abuse, and vulgarity in online memes. Through extensive experiments on four different language datasets, our results demonstrate that transformer-based VisualBERT outperforms traditional convolutional models such as ResNet in effectively capturing multimodal representations.

This work provides a comparative analysis between two multimodal architectures, highlighting the strengths of transformer-based fusion in handling both textual and visual information. Beyond the image and textual feature extraction methods employed in this study, several alternative strategies for feature extraction and multimodal fusion remain unexplored. Future research may focus on incorporating advanced vision-language models, attention-based fusion strategies, or contrastive learning approaches to further enhance performance in multilingual meme classification tasks.

Declaration on Generative AI

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

References

[1] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys (CSUR) 51 (2018). doi:https://doi.org/10.1145/3232676.

[2] J. Kansok-Dusche, C. Ballaschk, N. Krause, A. Zeißig, L. Seemann-Herz, S. Wachs, L. Bilz, A systematic review on hate speech among children and adolescents: Definitions, prevalence, and overlap with related phenomena. trauma, violence, & abuse, Trauma Violence Abuse 24 (2023) 2598–2615. doi:https://doi.org/10.1177/15248380221108070.

[3] 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, 2019, pp. 14–17. doi:https://doi.org/10.1145/3368567.33685.

[4] 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, 2019, p. 14–17. doi:10.1145/3368567.3368584.

[5] 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 (2026) 303–309. URL: https://www.sciencedirect.com/science/article/pii/S0167865525003150. doi:https://doi.org/10.1016/j.patrec.2025.09.004.

[6] K. Ghosh, N. K. Singh, J. Mahapatra, S. Saha, A. Senapati, U. Garain, Safespeech: a three-module pipeline for hate intensity mitigation of social media texts in indic languages, Social Network Analysis and Mining 14 (2024) 245. doi:https://doi.org/10.1007/s13278-024-01393-9.

[7] 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.

[8] 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.
[9] 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 (Eds.), Forum for Information Retrieval Evaluation (Working Notes) (FIRE 2025) December 17-20, Varanasi , India, CEUR-WS.org, 2025.
[10] 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.
[11] 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: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, Gandhinagar, India, December 13-17, 2021, volume 3159 of *CEUR Workshop Proceedings*, 2021, pp. 446–458. URL: https://ceur-ws.org/Vol-3159/T1-44.pdf.
[12] A. Saroj, S. Chanda, S. Pal, IRlab@IITV at SemEval-2020 task 12: Multilingual offensive language identification in social media using SVM, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 2012–2016. URL: https://aclanthology.org/2020.semeval-1.265/.
[13] S. Chanda, A. Dhaka, S. Pal, Crossing borders: Multilingual hate speech detection, in: FIRE (Working Notes), 2023, pp. 486–500. URL: https://ceur-ws.org/Vol-3681/T6-15.pdf.
[14] 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: FIRE (Working Notes), 2022, pp. 502–512. URL: https://ceur-ws.org/Vol-3395/T7-3.pdf.
[15] S. Chanda, A. Dhaka, S. Pal, Towards safer online spaces: Deep learning for hate speech detection in code-mixed social media conversations, in: Companion Publication of the 16th ACM Web Science Conference, Association for Computing Machinery, 2024, pp. 103–109. doi:10.1145/3630744.3663610.
[16] S. Chanda, S. Pal, Hate content identification in code-mixed social media data, in: Text and Social Media Analytics for Fake News and Hate Speech Detection, Chapman and Hall/CRC, 2024, pp. 225–247. doi:10.1201/9781003409519-13.
[17] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (1998) 2278–2324. doi:10.1109/5.726791.
[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017). URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
[19] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems 25 (2012). doi:https://doi.org/10.1145/3065386.
[20] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR abs/1409.1556 (2014). doi:https://api.semanticscholar.org/CorpusID:14124313.
[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions , in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9. doi:https://doi.ieeecomputersociety.org/10.1109/CVPR.2015.7298594.
[22] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 770–778. URL: https://api.semanticscholar.org/CorpusID:206594692.
[23] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers

for language understanding, in: North American Chapter of the Association for Computational Linguistics, 2019. URL: https://api.semanticscholar.org/CorpusID:52967399.

[24] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, Visualbert: A simple and performant baseline for vision and language, ArXiv abs/1908.03557 (2019). URL: https://api.semanticscholar.org/CorpusID:199528533.