Overview of the Multimodal Homophobia and Transphobia Meme Classification Shared Task

Kishore Kumar Ponnusamy[1], Bharathi Raja Chakravarthi[2] , Mahesh Susaladi[2], Junru Ren[2], Prasanna Kumar Kumaresan[2], Premjith B[3], Durairaj Thenmozhi[4], Ruba Priyadharshini[5], Subalalitha Chinnaudayar Navaneethakrishnan[6]

[1]Digital University Kerala, India,
[2]Data Science Institute, University of Galway, Ireland,
[3]Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India,
[4]Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu, India,
[5]Gandhigram Rural Institute - Deemed to be University, Tamil Nadu, India,
[6]SRM Institute of Science and Technology, Tamil Nadu, India

Correspondence: bharathi.raja@universityofgalway.ie

Abstract

This paper presents an overview of the Shared Task on detecting homophobia and transphobia in meme datasets across three languages: Hindi, English, and Chinese. With the rapid growth of internet users worldwide, memes have become a widely used medium for expressing humor, satire, and sarcasm on social media platforms. However, their increasing popularity has also facilitated the spread of hate, misinformation, and propaganda targeting specific communities. Hateful memes often attack individuals or groups based on attributes such as physical appearance, language, ethnicity, religion, or sexual orientation. Among those affected, the LGBTQ+ community is particularly vulnerable and frequently targeted on social media platforms. To address this issue, we organized a shared task that focuses on identifying homophobic and transphobic hate in memes. The task aims to encourage the development of automated systems capable of detecting such harmful content across multiple languages. Evaluation was conducted using Macro F1-score as the primary metric. The top performing system achieved a Macro F1-score of 0.8377 for English, 0.8081 for Hindi, and 0.7535 for Chinese, demonstrating promising results for multilingual hate detection in memes.

Disclaimer: This paper (including figures and examples) may contain offensive, hateful, or harmful language and imagery, including content targeting individuals based on sexual orientation or gender identity. All such material is presented strictly for research and educational purposes to support the development of automatic detection systems. The content does not reflect the views of the authors or the organizers. Reader discretion is advised.

1 Introduction

Homophobia and transphobia refer to negative comments, including hatred, discomfort, or prejudice, aimed at lesbian, gay, bisexual, and transgender individuals (Chakravarthi, 2024). Expressions of homophobic or transphobic feelings often involve foul language and lead to hate speech targeting these groups, this type of content is becoming more prevalent on the internet. The presence of homophobia and transphobia on social media poses a serious challenge, as it spreads harmful content that undermines equality, diversity, and social inclusion (Chakravarthi, 2024). Such content is commonly present online in the form of memes, images, videos, and comments on social media.

According to the Cambridge English Dictionary, a meme is "an idea, joke, image, video, etc. that is spread very quickly on the internet" (Cambridge English Dictionary, n.d.). Over the years, memes have become a popular way to express humor, satire, and sarcasm online. Besides mainstream social media platforms like Facebook, Instagram, and Reddit, many platforms have emerged specifically for creating and sharing memes, such as Memechat, Pinterest, and others (Joshi et al., 2024).

Initially, memes were predominantly shared in English; however, in recent years, their use has expanded significantly across regional languages in India, particularly Hindi. This shift is unsurprising given that Hindi is one of the most widely spoken languages globally, with over 600 million speakers. After English and Chinese, Hindi is the third most spoken language in the world (Eberhard et al., 2021). Mostly used in India and the Indian subcontinent, it is one of the official languages of India and is also recognized as a protected language

Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 141–149

in Fiji, Nepal, South Africa, and the United Arab Emirates (Wikipedia contributors, 2025).

English, with approximately 1.4–1.5 billion speakers worldwide (including native and non-native speakers), is the most widely spoken language globally. It serves as the primary language of international communication, academia, business, and digital media, making it the dominant language in the early evolution of internet culture and meme dissemination. Similarly, Chinese particularly Mandarin Chinese has over 1.1 billion speakers, primarily concentrated in China and across Chinese-speaking communities worldwide. As the official language of the People's Republic of China and Taiwan, and one of the six official languages of the United Nations, Mandarin has played a significant role in shaping internet culture within Chinese digital ecosystems. The growth of social media platforms in China has further facilitated the rapid creation and spread of memes in Chinese, contributing to a distinct and highly dynamic online meme culture.

Given the large number of social media users, hate speech targeting marginalized communities is widely observed online (Chakravarthi et al., 2021a) and has increasingly taken multi-modal forms such as memes. Prior studies have shown that such hateful content targeting the LGBTQ+ community contributes to hostile online environments, making social media platforms unsafe and socially isolating for affected individuals (Sánchez-Sánchez et al., 2024). Despite the large population of English,Hindi and Chinese speaking social media users and the existence of hateful content in meme formats, studies specifically focused on identifying homophobia and transphobia in Hindi memes remain extremely limited.

2 Related Work

Memes communicate meaning through the interaction of images and text, making them fundamentally multimodal. Determining whether a meme is hateful therefore requires analyzing both components together rather than in isolation. Prior research demonstrates that unimodal systems perform poorly in this setting: text-only models achieve roughly 65 accuracy, and image-only systems perform even worse. In contrast, multimodal approaches that jointly model visual and textual information improve performance to around 70–75 (Kiela et al., 2020; Das et al., 2020). This improvement stems from what (Das et al., 2020) describe as "benign confounders," where individually harmless text and images produce hateful meaning when combined. To address this challenge, Vision-Language Pretrained Models (VL-PTMs) such as OSCAR, CLIP, and BERT integrate visual and textual representations through cross-modal fusion, enabling more effective meme understanding (Chen and Pan, 2022).

Even though multi modal meme detection has progressed significantly, research specifically targeting detection of homophobia and transphobia remains limited. Existing LGBTQ+ hate speech datasets, such as the one introduced by (Chakravarthi et al., 2021b), focus primarily on textual comments not multi modal memes. The LT-EDI Shared Task (Chakravarthi et al., 2022) expanded LGBTQ+ hate detection across five languages such as English, Spanish, Tamil, Hindi, and Malayalam. Meanwhile, multilingual and code-mixed hate speech research has grown, with studies on Hinglish memes (Rajput et al., 2022), foundational Hindi-English datasets (Bohra et al., 2018; Singh and Lefever, 2020), and the Chinese harmful meme dataset TOXICNMM (Lu et al., 2024) demonstrating the complexity of culturally grounded, multilingual meme interpretation. Cross-lingual approaches using models such as XLM-RoBERTa further show that multilingual systems can achieve strong results even with limited labeled data (Mnassri et al., 2024; Dinarta and Wicaksana, 2025).

The current research landscape reveals a clear gap at the intersection of three areas: multimodal meme detection for homophobia and transphobia, multilingual modeling, and LGBTQ+-targeted hate speech analysis. While each of these areas has been studied independently, their integration remains largely underexplored. To address this gap, we organized the Overview of Multimodal Homophobia and Transphobia Meme Classification Shared Task. Through this shared task, we introduce a novel multilingual multimodal dataset in English, Hindi, and Chinese to support research on detecting homophobic and Transphobic memes.

3 Task Description

The Multimodal Homophobia and Transphobia Meme Classification Shared Task aims to automatically detect and categorize harmful memes targeting LGBTQ+ communities. Participants were

provided with annotated training and test datasets to develop and evaluate their systems. Each meme must be classified into one of three categories: Homophobia, Transphobia, or Non-LGBT, using both textual and visual information incorporated in the meme.

The task incorporates memes in three languages English, Hindi, and Chinese representing diverse cultural contexts. This multilingual setup enables the development and evaluation of systems capable of detecting harmful content across multiple languages and cultural settings. It also promotes research on cross-lingual and cross-cultural multimodal hate detection against LGBTQ+ Communities, highlighting the challenges involved in identifying harmful narratives across different online platforms.

4 Dataset Description

The dataset for the Homophobia and Transphobia Meme Detection task consists of multimodal memes collected from social media platforms. Each meme contains both visual content (image) and associated textual information, either provided as captions or embedded within the image itself. The dataset captures diverse expressions of homophobia, transphobia, and neutral or non-anti-LGBT content, reflecting real-world online discourse and harmful meme culture across multiple languages, including English, Hindi, and Chinese. This diversity makes the dataset well-suited for multimodal hate meme classification.

The dataset was divided into training and testing sets for each language track. The training sets were provided with class labels, while the test sets were released without labels for evaluation. Each meme instance is annotated into one of three categories: Homophobia, Transphobia, or Non-LGBT. The data distribution and class distribution of training and test sets for all language tracks are presented in Table 1.

Table 1: Dataset Statistics

Class	English		Hindi		Chinese	
	Train	Test	Train	Test	Train	Test
Homophobic	160	40	366	64	705	176
Transphobic	160	41	274	96	55	14
Non-LGBT	240	60	158	40	196	49
Total	560	141	798	200	956	239

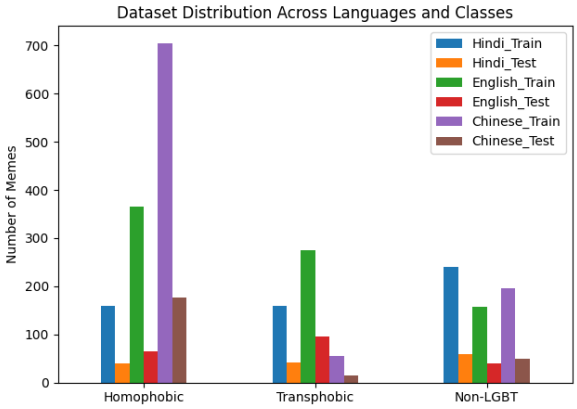

Figure 1: Dataset Distribution

5 Dataset Construction

5.1 Data Collection From Social Media Platforms

On social media, hashtags act as labels that group large volumes of related content such as memes, images, and short videos, making topics easier to discover and spread across online communities (Caleffi, 2015). In the context of transphobic and homophobic memes, hashtags are often derived from stigmatizing or derogatory terms, which allows such content to be easily located and circulated. Platforms like Instagram, X (formerly Twitter), and Facebook enable meme collection through these hashtags, while Reddit communities (subreddits) organize discussions and posts around similar themes. Additionally, Pinterest provides searchable meme templates that can be reused or adapted to create and analyze meme content across platforms.

5.2 Data Augmentation

To generate a multimodal meme, both visual content (an image template) and textual content (a caption) are required. Several online platforms support meme creation; among them, Imgflip is a widely used service that provides a large collection of popular meme templates. Imgflip allows users to generate memes either through its graphical user interface (GUI) or via its meme generation API. The API enables automated and scalable meme creation, making it particularly useful for generating large numbers of memes for dataset expansion and data augmentation.

For Hindi, we already had access to an existing dataset containing derogatory comments and hate speech, which we used as captions for meme generation. We carefully examined the human intent

and emotional context behind each derogatory remark, selected an appropriate meme template that matched the sentiment, and then added the caption to the template to generate the final meme.

Figure 2: English Meme Examples

Figure 3: Hindi Meme Examples

Figure 4: Chinese Meme Examples

6 Methodology

A total of participating teams created systems for the Homophobia and Transphobia Meme Classification shared task, which took place in English, Hindi, and Chinese. You had to sort memes into three groups: Homophobia, Transphobia, and Non-Anti-LGBT. Because memes can be in more than one form, most systems combined text and images, often using OCR pipelines to pull out text that was hidden in images.

The **MemeScouts** (Bueno et al., 2026) team adopted a Prompted Weak Supervision strategy. A Vision–Language Model (VLM) was prompted with 89 structured yes/no or either/or questions about each meme. The generated responses were treated as structured features and used to train a Random Forest classifier. We used feature selection based on the validation F1-score to get rid of extra or noisy attributes. This method put a lot of emphasis on how easy it is to understand and how well it can adapt to changing meme trends.

The **BiasBreakers** team developed a multimodal pipeline combining visual and textual signals. Text was extracted using EasyOCR and a Vision–Language Model, and a perplexity-based scoring mechanism selected the most coherent text representation. Both image and selected text were encoded using a pretrained CLIP model into a shared embedding space. The concatenated embeddings were passed through a lightweight neural classifier optimized using cross-entropy loss and AdamW.

The **SigJBS** (Sinha et al., 2026) team implemented an OCR-aware multimodal framework. Meme text was extracted and normalized before being combined with image inputs in a LoRA-fine-tuned Qwen2-VL-2B vision–language model. A zero-shot CLIP baseline was also evaluated. Parameter-efficient fine-tuning enabled adaptation with limited computational resources while accounting for class imbalance using macro-F1-based validation.

The **susmitha** (Jaishri et al., 2026) team proposed a gated multimodal fusion architecture. XLM-RoBERTa (base) was used for multilingual textual embeddings, and CLIP-ViT-B/32 was used for visual feature extraction. OCR preprocessing using Tesseract supplemented embedded text detection. A learnable gating mechanism dynamically adjusted the weight of each modality's contribution, and weighted cross-entropy loss was used to fix bias and imbalance.

The **SAJI** (Bandyopadhyay et al., 2026) team employed a zero-shot sequential transformer approach. Qwen 2.5 VL was used for extracting and analyzing meme content, and then they used Llama 3 and Mistral large language models to sort it using prompt-based inference. The system produced single-word outputs without any supervision or fine-tuning.

The **EthosAI** team adopted an element-wise fusion strategy. EfficientNet-B0 extracted visual features, and paraphrase-multilingual-MiniLM-L12-v2 generated textual embeddings. Both feature vectors were projected to the same dimension and fused element-wise before classification, emphasizing computational efficiency.

The **DLRG** team treated the problem as a pure image classification task. EfficientNet-B3 was used as a pretrained backbone for feature extraction, fol-

lowed by a linear classification head for predicting the three classes. This approach excluded textual information.

The **CuriousVectors** (Kushwaha et al., 2026) team extracted meme text using Tesseract OCR and fine-tuned RoBERTa for text encoding and ConvNeXT for image encoding. Both representations were projected to 256-dimensional vectors and combined. An XGBoost classifier was trained on fused features for final prediction.

The **MemeSentinel** team build a CLIP-based multimodal architecture with a gated fusion module. Image and textual embeddings were combined adaptively, and test-time augmentation (TTA) was used during inference to make the system more robust. The system was evaluated separately for English, Hindi, and Chinese tracks.

Across all submissions, multimodal architectures were dominant. OCR-based preprocessing and CLIP or VLM-based encoders were frequently adopted. Fusion mechanisms included concatenation, element-wise operations, and gated adaptive weighting. Some teams looked into zero-shot and prompt-based methods, while others worked on supervised multimodal fine-tuning.multimodal fine-tuning.

7 Results and Discussion

The submitted systems were evaluated using the Macro-averaged F1-score (MF1) as the primary metric. We also used Accuracy (ACC), Macro Precision (MP), Macro Recall (MR), Weighted Precision (WP), Weighted Recall (WR), and Weighted F1 (WF1). We chose Macro-F1 to ensure that all three classes performed equally well. This was especially important because there could be an imbalance between memes that are homophobic, transphobic, and not anti-LGBT.

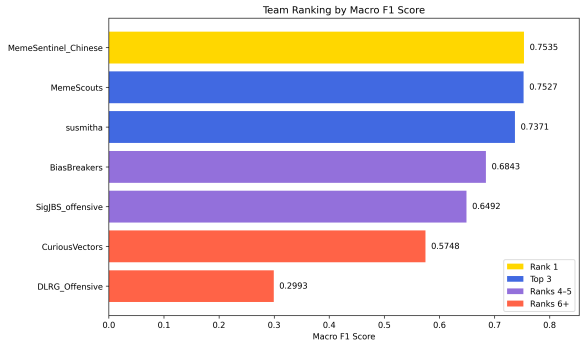

Figure 5: Macro F1 Ranking Chart For Chinese Memes

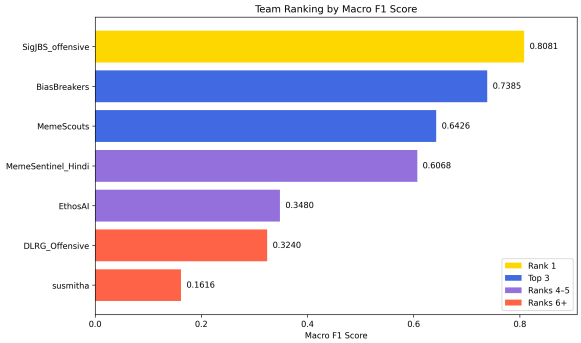

Figure 6: Macro F1 Ranking Chart For Hindi Memes

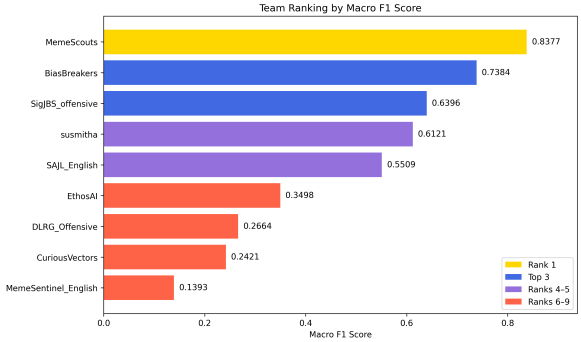

Figure 7: Macro F1 Ranking Chart For English Memes

7.1 English Track

In the English track, **MemeScouts** secured first rank with a Macro-F1 score of 0.8377, achieving the highest balanced performance across categories. The high macro precision (0.8405) and macro recall (0.8393) show that the model can tell the difference between classes well. The Prompted Weak Supervision method showed that structured VLM-derived features can effectively capture multimodal hate signals.

The **BiasBreakers** team ranked second with a Macro-F1 of 0.7384, followed by **SigJBS** with 0.6396. These systems leveraged CLIP-based and LoRA-fine-tuned VLM architectures, confirming the effectiveness of OCR-aware multimodal modeling.

Mid-ranked systems like **susmitha** (0.6121) and **SAJI**(0.5509) showed moderate performance. The gated multimodal fusion approach outperformed zero-shot prompt-based inference, suggesting the benefit of supervised fine-tuning.

DLRG (0.2664) and **MemeSentinel** (0.1393) are two examples of systems that don't work well with unimodal modeling or multimodal fusion in the English dataset.

Table 2: Leaderboard of Participating Systems Ranked by Macro-F1 (Homophobia and Transphobia Meme Classification)

S.No	Team	Run	Acc	MP	MR	**Macro F1**	WP	WR	WF1	Rank
					ENGLISH					
1	MemeScouts	Run2	0.8369	0.8405	0.8393	**0.8377**	0.8372	0.8369	0.8352	1
2	BiasBreakers	Run1	0.7376	0.7495	0.7701	**0.7384**	0.7656	0.7376	0.7277	2
3	SigJBS_offensive	Run1	0.6525	0.6654	0.6300	**0.6396**	0.6605	0.6525	0.6481	3
4	susmitha	Run1	0.6454	0.7171	0.6079	**0.6121**	0.7114	0.6454	0.6323	4
5	SAJL_English	Run1	0.5887	0.6481	0.5444	**0.5509**	0.6326	0.5887	0.5661	5
6	EthosAI	Run2	0.4255	0.3777	0.3742	**0.3498**	0.4078	0.4255	0.3926	6
7	DLRG_Offensive	Run1	0.3333	0.2633	0.3026	**0.2664**	0.2803	0.3333	0.2909	7
8	CuriousVectors	Run1	0.3475	0.3144	0.2883	**0.2421**	0.3268	0.3475	0.2791	8
9	MemeSentinel_English	Run1	0.2766	0.1500	0.1300	**0.1393**	0.3191	0.2766	0.2764	9
					HINDI					
1	SigJBS_offensive	Run1	0.8400	0.8217	0.7986	**0.8081**	0.8398	0.8400	0.8878	1
2	BiasBreakers	Run1	0.7800	0.7392	0.7479	**0.7385**	0.7776	0.7800	0.7628	2
3	MemeScouts	Run2	0.6550	0.6397	0.6597	**0.6426**	0.6780	0.6550	0.6605	3
4	MemeSentinel_Hindi	Run2	0.6150	0.6071	0.6330	**0.6068**	0.6500	0.6150	0.6199	4
5	EthosAI	Run2	0.4050	0.4201	0.3906	**0.3480**	0.4662	0.4050	0.3747	5
6	DLRG_Offensive	Run1	0.3600	0.3420	0.3382	**0.3240**	0.3771	0.3600	0.3521	6
7	susmitha	Run1	0.3200	0.1067	0.3333	**0.1616**	0.1024	0.3200	0.1552	7
					CHINESE					
1	MemeSentinel_Chinese	Run1	0.8636	0.8298	0.7116	**0.7535**	0.8526	0.8636	0.8488	1
2	MemeScouts	Run1	0.8243	0.8002	0.7350	**0.7527**	0.8396	0.8243	0.8276	2
3	susmitha	Run1	0.8452	0.7644	0.7225	**0.7371**	0.8471	0.8452	0.8447	3
4	BiasBreakers	Run1	0.8075	0.8242	0.6247	**0.6843**	0.8061	0.8075	0.7980	4
5	SigJBS_offensive	Run1	0.8285	0.6807	0.6296	**0.6492**	0.8182	0.8285	0.8220	5
6	CuriousVectors	Run1	0.7500	0.5912	0.5874	**0.5748**	0.7681	0.7500	0.7522	6
7	DLRG_Offensive	Run1	0.5858	0.2946	0.3044	**0.2993**	0.5603	0.5858	0.5727	7

Acc: Accuracy; MP: Macro-Precision; MR: Macro-Recall; WP: Weighted-Precision; WR: Weighted-Recall; WF1: Weighted F1.

7.2 Hindi Track

In the Hindi track, **SigJBS** achieved first place with a Macro-F1 of 0.8081, demonstrating strong OCR-aware VLM fine-tuning capability. The system kept a good balance between high macro precision and recall.

BiasBreakers ranked second (0.7385), while **MemeScouts** secured third place (0.6426). Unlike the English track, the supervised multimodal LoRA-based model outperformed the prompted feature-based method, indicating that language-specific adaptation plays a critical role in Hindi meme classification.

Systems such as **MemeSentinel** (0.6068) showed competitive performance. But lightweight fusion methods like **EthosAI** (0.3480) and unimodal image classification methods like **DLRG** (0.3240) achieved comparatively lower Macro-F1 scores. The significant drop for some teams indicates higher linguistic complexity or OCR challenges in Hindi memes.

7.3 Chinese Track

In the Chinese track, **MemeSentinel** came in first with a Macro-F1 score of 0.7535. This shows that gated multimodal CLIP-based fusion works well when combined with test-time augmentation.

MemeScouts came in second with 0.7527, showing consistent cross-lingual adaptability of the prompted weak supervision strategy. **susmitha** ranked third (0.7371), indicating the strength of gated multimodal fusion in multilingual contexts.

Interestingly, performance gaps were narrower in Chinese compared to English. CLIP-based systems generalized effectively, while unimodal systems such as **DLRG** (0.2993) again showed lower performance.

8 Conclusion

The Multimodal Homophobia and Transphobia Meme Classification Shared Task tackled a important and underexplored challenge at the convergence of multimodal hate detection, multilingual modeling, and LGBTQ+-targeted content analysis. The task established a structured benchmark for

Table 3: Summary of Participating Systems for Homophobia and Transphobia Meme Detection (✓= reported, × = not reported).

Team	Model(s) Reported	OCR	VLM	Fusion	Img	Txt	Zero	LLM	Aug
BiasBreakers	CLIP + Neural Classifier	✓	✓	Concat	✓	✓	✗	✗	✗
MemeSentinel	CLIP + Gated Fusion	✓	✓	Gated	✓	✓	✗	✗	✓
MemeScouts	VLM Prompting + RF	✗	✓	Feature-based	✗	✓	✓	✓	✗
EthosAI	EfficientNet-B0 + MiniLM	✗	✗	Element-wise	✓	✓	✗	✗	✗
CuriousVectors	RoBERTa + ConvNeXT + XG-Boost	✓	✗	Projection + ML	✓	✓	✗	✗	✗
SAJI	Qwen-VL + LLaMA + Mistral	✓	✓	Sequential	✗	✓	✓	✓	✗
SigJBS	Qwen2-VL + LoRA	✓	✓	Joint VLM	✓	✓	✓	✓	✗
susmitha	XLM-R + CLIP-ViT	✓	✓	Gated	✓	✓	✗	✗	✗
DLRG	EfficientNet-B3	✗	✗	Image-only	✓	✗	✗	✗	✗

OCR: Optical Character Recognition; **VLM**: Vision-Language encoder; **Fusion**: Multimodal fusion strategy; **Img**: Dedicated image encoder; **Txt**: Dedicated text encoder; **Zero**: Zero-shot inference; **LLM**: LLM prompting or LoRA fine-tuning; **Aug**: Data augmentation or test-time augmentation.

assessing automatic systems intended to identify harmful meme content aimed at LGBTQ+ communities by introducing an innovative multilingual dataset that includes English, Hindi, and Chinese memes. The results show that multimodal approaches always do better than unimodal baselines. It demonstrates that it is essential to model both visual and textual signals together when trying to understand memes. Across all language tracks, systems that used OCR-aware pipelines, CLIP-based encoders, and vision-language pretrained models did the best, with Macro-F1 scores of 0.8377 in English, 0.8081 in Hindi, and 0.7535 in Chinese. The differences in performance between languages show how linguistic complexity, OCR reliability, and cultural meme dynamics affect how well a model works. While supervised multimodal fine-tuning and gated fusion strategies proved particularly robust, the comparatively lower performance of image-only systems underscores the importance of integrated multimodal reasoning. Overall, the shared task establishes a solid groundwork for future research in multilingual multimodal hate detection. It also emphasizes the need for culturally grounded modeling, better cross-lingual generalization, and AI systems that are ethically responsible to make online spaces safer for marginalized communities.

9 Limitations

The shared task sets a useful standard for finding memes that are homophobic and transphobic in multiple languages, but there are still some problems. One limitation is that the meme content comes from social media sites, which may have their own biases and not show all types of harmful content equally. Memes often combine images, embedded text, sarcasm, and cultural references, making them inherently ambiguous and difficult to interpret even for human annotators. The task is even harder because English, Hindi, and Chinese all have different languages, slang, and cultural contexts. This could make it harder for models trained on the dataset to generalize. Another limitation comes from the multimodal nature of memes, where harmful meaning may be communicated through nuanced interactions between visual components and textual indicators, which many models still struggle to capture effectively. In addition, the dataset size is relatively limited compared to large-scale multimodal benchmarks, which may restrict the ability of deep learning models to learn robust representations across languages and cultural contexts. These problems show that we need bigger multilingual multimodal datasets, better ways to annotate across cultures, and more advanced multimodal reasoning models to find online hate that targets LGBTQ+ communities that is not obvious and depends on the situation.

Ethical Consideration

This shared task dataset contains meme content that may have language and images that are homophobic, transphobic, or otherwise offensive to LGBTQ+ people and groups. Such material is included strictly for research and educational purposes and does not reflect the views of the authors or organizers. The memes were collected from publicly available online sources, and steps were taken to remove or avoid including any information that

could be used to identify a person in order to protect their privacy. Only the minimal data necessary for research and system evaluation was released. The annotation process was carried out by trained annotators following clear labeling guidelines designed to ensure responsible handling of sensitive content. Despite these precautions, the dataset may still contain biases present in online discourse or cultural interpretations of memes. Therefore, the dataset should be used responsibly and with awareness of potential risks, including bias amplification or misuse in automated moderation systems. We support open research practices and suggest that people keep an eye on systems trained on this dataset to help with fair and ethical content moderation.

Acknowledgement

The authors, Bharathi Raja Chakravarthi was funded by a research grant from Research Ireland under grant number SFI/12/RC/2289_P2 (Insight_2) and Prasanna Kumar Kumaresan funded by Research Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223.

References

Jishnu Bandyopadhyay, Saloni Kushwaha, Deepawali Sharma, and Aakash Singh. 2026. Saji_english@lt-edi 2026: Detection of homophobia and transphobia in internet memes using zero-shot learning. In *Proceedings of Sixth Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of hindi-english code-mixed social media text for hate speech detection. In *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*, pages 36–41.

Ivo Bueno, Lea Hirlimann, and Enkelejda Kasneci. 2026. Memescouts@lt-edi 2026: Asking the right questions - prompted weak supervision for meme hate speech detection. In *Proceedings of Sixth Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Paola-Maria Caleffi. 2015. The 'hashtag': A new word or a new rule? *SKASE Journal of Theoretical Linguistics*, 12(2):46–69.

Cambridge English Dictionary. n.d. Meme. https://dictionary.cambridge.org/dictionary/english/meme. Retrieved on 2025-12-17.

Bharathi Raja Chakravarthi. 2024. Detection of homophobia and transphobia in youtube comments. *International Journal of Data Science and Analytics*, 18(1):49–68.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. 2021a. Homophobia and transphobia detection in social media comments. In *Proceedings of the 15th International Conference on Semantic Evaluation (SemEval-2021)*, pages 739–744. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, Subalalitha Cn, John Philip McCrae, Miguel Ángel García, Salud María Jiménez-Zafra, Rafael Valencia-García, Prasanna Kumaresan, Rahul Ponnusamy, and 1 others. 2022. Overview of the shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the second workshop on language technology for equality, diversity and inclusion*, pages 378–388.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kayalvizhi Sampath, Durairaj Thenmozhi, Sathiyaraj Thangasamy, Rajendran Nallathambi, and John Phillip McCrae. 2021b. Dataset for identification of homophobia and transophobia in multilingual youtube comments. *arXiv preprint arXiv:2109.00227*.

Yuyang Chen and Feng Pan. 2022. Multimodal detection of hateful memes by applying a vision-language pre-training model. *Plos one*, 17(9):e0274300.

Abhishek Das, Japsimar Singh Wahi, and Siyao Li. 2020. Detecting hate speech in multi-modal memes. *arXiv preprint arXiv:2012.14891*.

Farrel Dinarta and Arya Wicaksana. 2025. Enhanced hate speech detection in indonesian-english code-mixed texts using xlm-roberta. *Informatica*, 49(21).

David M Eberhard, Gary F Simons, and Charles D Fennig. 2021. What are the top 200 most spoken languages. *Ethnologue: Languages of the world*.

N.P.Susmitha Jaishri, Kogilavani Shanmugavadivel, Malliga Subramaniyan, and Mouleeshuwarapprabu R. 2026. Susmitha@lt-edi 2026: Detecting lgbtq+ phobia in multilingual memes via joint representation. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Saurav Joshi, Filip Ilievski, and Luca Luceri. 2024. Contextualizing internet memes across social media platforms. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1831–1840.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and

Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Saloni Kushwaha, Jishnu Bandyopadhyay, Deepawali Sharma, and Aakash Singh. 2026. Curiousvectors@lt-edi 2026: Detection of homophobic and transphobic memes on social media using a hybrid multimodal approach. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Junyu Lu, Bo Xu, Xiaokun Zhang, Hongbo Wang, Haohao Zhu, Dongyu Zhang, Liang Yang, and Hongfei Lin. 2024. Towards comprehensive detection of chinese harmful memes. *Advances in Neural Information Processing Systems*, 37:13302–13320.

Khouloud Mnassri, Reza Farahbakhsh, and Noel Crespi. 2024. Multilingual hate speech detection: a semi-supervised generative adversarial approach. *Entropy*, 26(4):344.

Kshitij Rajput, Raghav Kapoor, Kaushal Rai, and Preeti Kaur. 2022. Hate me not: detecting hate inducing memes in code switched languages. *arXiv preprint arXiv:2204.11356*.

Ana M Sánchez-Sánchez, David Ruiz-Muñoz, and Francisca J Sánchez-Sánchez. 2024. Mapping homophobia and transphobia on social media. *Sexuality Research and Social Policy*, 21(1):210–226.

Pranaydeep Singh and Els Lefever. 2020. Sentiment analysis for hinglish code-mixed tweets by means of cross-lingual word embeddings. In *Proceedings of the 4th Workshop on Computational Approaches to Code Switching*, pages 45–51.

Gaurangi Sinha, Rajarajeswari Palacharla, and Manoj Balaji Jagadeeshan. 2026. Sigjbs@lt-edi 2026: Multimodal homophobia and transphobia meme classification. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Wikipedia contributors. 2025. Hindi. Accessed: 17 Dec 2025.