CuriousVectors@LT-EDI 2026: Detection of Homophobic and Transphobic Memes on Social Media Using a Hybrid Multimodal Approach

Saloni Kushwaha[1] **Jishnu Bandyopadhyay**[1] **Deepawali Sharma**[2] **Aakash Singh**[1]
[1]Department of Computer Science, University of Delhi, India
[2]School of Computer Science Engineering and Technology, Bennett University, Noida, India
{salonimsc24,jishnumsc24, asingh}@cs.du.ac.in
deepawali21@bhu.ac.in

Abstract

The rapid growth of social media has also led to a rise in abusive and harmful content, which negatively affects the online environment for users. The frequent use of offensive language and hate speech contributes to making these platforms increasingly hostile. In particular, homophobic and transphobic remarks target members of the LGBT+ community. Detecting such comments is therefore essential so that they can be flagged promptly and appropriate warnings can be given to users involved in such behaviour. The problem becomes more serious when such content appears in other forms of communication used by younger generations, such as memes. This work tries to address this issue. We propose a method to detect such content using the meme dataset from the LT-EDI 2026 challenge and secured 8th rank for English and 6th rank for Chinese language dataset in the shared task. Our approach uses a multimodal technique that processes both image and text information. The dataset has limited data, which creates a challenge. To handle this, we pre–fine-tune the models on a similar dataset called PrideMM. The proposed multimodal approach achieved Macro F1-scores of 0.24 and 0.57 for English and Chinese memes respectively.

1 Introduction

Over the past decade,social media has become one of the major parts of the modern world. As of early 2026 , there are over 5.66 billion active users worldwide,representing nearly 2 in 3 people on earth who use social media[1]. Social media has provided a new way of life to everyone where they can put forward their views among billions of people without explicit physical presence. These platforms have been used to discuss the news about what is happening around the world, social moments,entertainment etc(Singh et al., 2026). Apart from all the positive impacts it provides,people use it to spread hate against certain communities like women,specially abled people,religion as well as LGBT+ (Sharma et al., 2023). These hateful comments are not limited to text only but also images,videos,which require special computational tools (Singh et al., 2025) ,(Singh et al., 2024).

Homophobia/ Transphobia is a serious abuse that can take the shape of physical violence such as murder, beating, rape, molestation, privacy violation (Chakravarthi, 2024). Studies indicate that approximately 93% of transgender individuals experience online harassment, compared to 70% of cisgender individuals (Chakravarthi, 2024). These statistics clearly show the seriousness of the problem and the urgent need for effective solutions. Although several studies (Mossie and Wang, 2020), (Arcila-Calderón et al., 2022) have been conducted on detecting hate speech against vulnerable communities. There is limited studies that focused on detection of hate from memes, which often express discrimination through sarcasm, hidden messages, or symbolic images. This study attempts to bridge this gap by detecting homophobia and transphobia using the LT-EDI 2026 dataset.

The major contribution of this work is the development of an effective multimodal approach that is able to extract meaningful and connected information from both text and images. For text learning, RoBERTa and ChineseBERT models were used, while ConvNeXt was applied for learning image features, allowing the system to handle different types of data efficiently. To further improve the understanding of anti-LGBT+ content, the model for the English language was fine-tuned on a related publicly available dataset called PrideMM. This additional training helped the models better adapt to similar real-world content and improved their overall ability to analyze sensitive multimodal

[1]https://datareportal.com/reports/digital-2026-two-in-three-people-use-social-media

Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 155–160

data.

The rest of the paper is organized as follows. Section 2 reviews related work on hate speech detection targeting various communities, including women, religious groups, and LGBT+ individuals. Section 3 describes the dataset used in this study. Section 4 details the proposed methodology, computational models, and hyperparameter tuning strategies. Section 5 presents experimental results and analysis. Finally, Section 6 concludes the paper and discusses potential directions for future research.

2 Related Work

With the increasing use of social media, many studies have examined the problem of hateful comments online (Gupta et al., 2023). Earlier, most of this harmful content appeared mainly in text form. However, over time it has started appearing in many other formats such as images (Lee et al., 2024), videos (Wang et al., 2024), and mixed forms like memes that combine text and visuals (Hermida and Santos, 2023). Because of this shift, detecting harmful content has become more challenging. Researchers have explored several types of targeted hate, including misogyny detection (Basile et al., 2019)(Shushkevich and Cardiff, 2019), binary classification of general hate speech (Gandhi et al., 2024), cyberbullying (Rosa et al., 2019), and religious hate (Sharma et al., 2024). As the amount of such content continues to grow, analysing it requires significant computational resources. The study of online hate is also no longer limited to Natural Language Processing alone. Work in this area now involves different approaches using Machine Learning, Deep Learning, and systems based on Large Language Models, which have helped make research in this field more advanced and robust (Channon and Mathieson, 2025a).

Several studies have highlighted the need to detect harmful content targeting the LGBT+ community (Sharma et al., 2023), (Channon and Mathieson, 2025b). However, most of this work has focused on a single modality, mainly text-based analysis. Research that examines such harmful content in multimodal formats remains relatively limited. The LT-EDI Workshop has created valuable opportunities to explore these challenges further (Ponnusamy et al., 2026). Over the years, this platform has introduced tasks on different forms of targeted hate speech, including Dravidian hate speech detection (Roy et al., 2022), multimodal hateful meme detection (Shah et al., 2024), and misogyny detection (Rahali et al., 2021). In the current edition, the homophobia and transphobia detection task has been extended to memes, where the content includes both text and images. This makes the problem inherently multimodal and encourages research that can analyse both visual and textual information together.

3 Dataset Description

The dataset used in this study was provided by the LT-EDI 2026 shared task (Ponnusamy et al., 2026). Two datasets were used in the proposed method. The PrideMM dataset was used for the first-stage of fine-tuning the multimodal architecture for the English language before classifying the dataset provided through the Shared Task of LT-EDI 2026 platform.

3.1 PrideMM Dataset

The PrideMM dataset contains data collected from several social media platforms such as Facebook, Twitter, and Reddit related to LGBT+ community (Shah et al., 2024). For hate speech detection, binary labels of hate and non-hate are provided and were used to fine-tune the multimodal architecture to reduce bias in the target task. Since the LT-EDI 2026 Shared Task 2 dataset is relatively small, the model was first fine-tuned on the larger PrideMM dataset to improve classification performance for the shared task for the English language.
Table 1 shows the data distribution of PrideMM dataset.

Table 1: Dataset distribution of PrideMM

Dataset	Hate	Non-Hate	Total
PrideMM	2581	2482	5063

3.2 Homophobia and Transphobia Meme Classification dataset (LT-EDI)

The LT-EDI 2026 shared task on homophobia and transphobia meme classification dataset (Ponnusamy et al., 2026) was available in three languages: English, Hindi, and Chinese. It was divided into two subsets, namely training and testing data. The dataset contains three classes: Homophobia, Transphobia, and Non-Anti-LGBT. Table 2 presents the distribution of the dataset across these classes.

Table 2: Dataset distribution across languages and train/test splits.

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_Anti_LGBT	240	60	158	40	196	49
Total	**560**	**141**	**798**	**200**	**956**	**239**

4 Methodology

This section describes the proposed framework for meme classification on the LT-EDI dataset. The multimodal architecture uses a ConvNeXt model for image-based learning and a RoBERTa model for text-based learning. Features from both modalities are combined using concatenation and then passed to an XGBoost classifier for the English language. For Chinese data, ConvNeXt is used with Chinese-BERT. This design highlights the importance of using both text and image information to understand the hidden interconnection often present in memes. The proposed method focuses on capturing sentiment from text and images together, which is difficult to achieve using only text-based or image-based models alone.

4.1 Data Preprocessing and Balancing

Images are processed using OpenCV, and each image is resized to 224 × 224 pixels to make it suitable for deep learning models. A linear intensity transformation is then applied with $\alpha = 1.2$ for contrast adjustment and $\beta = 20$ for brightness enhancement to improve visual features. Text data for both languages is prepared by extracting text from images using OCR and removing special characters, URLs, stopwords, and numbers. For the Chinese dataset, SMOTE is used to balance the data, and the Jieba tokenizer is applied for Chinese text tokenization.

4.2 Model Architecture

Experiments were first conducted on the LT-EDI dataset using several image-based models, including ResNet-50, DenseNet-121, Inception V3, and ConvNeXt. Among these, ConvNeXt achieved the best performance for both English and Chinese data, with macro F1 scores of 0.78 and 0.61 respectively. For English OCR text, multiple text-based models such as BERT, XLM- RoBERTa, RoBERTa with LoRA, RoBERTa, and DeBERTa were evaluated, and RoBERTa showed the best results with a macro F1 score of 0.68 For Chinese OCR text, models including BERT, ChineseBERT, and mBERT were tested, and ChineseBERT achieved the highest macro F1 score of 0.47. Based on these results, ConvNeXt, RoBERTa, and ChineseBERT were selected as the best-performing individual models for meme classification. Figure 1 shows the framework of proposed multimodal.

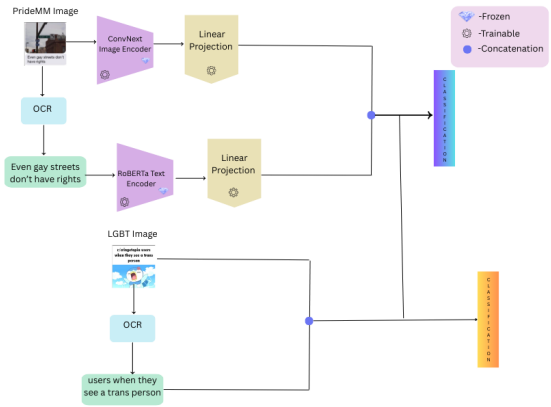

Figure 1: Proposed multimodal framework.

4.3 Architecture for English Data

For English data, two-stage fine tuning approach was used. The PrideMM dataset was used for the first-stage fine-tuning of the multimodal architecture. ConvNeXt was selected for image learning and RoBERTa for text learning, as they performed better than other tested models. The architecture uses the last three unfrozen layers of ConvNeXt and the last two layers of RoBERTa, followed by a linear projection layer (which is a single layer Artificial Neural Network) for each model. The projected features are concatenated to form a single multimodal representation, which is then passed to a single-layer perceptron for classification. In the second-stage fine-tuning phase on the LT- EDI dataset, the classifier head was removed and model weights were saved. The extracted features from LT-EDI Shared Task 2 data were then classified using Random Forest and XGBoost, where XGBoost achieved the best performance with a macro F1 score of 0.2421.

For the second-stage fine-tuning on the data provided by LT-EDI shared task, we have first dropped the classifier head of multimodal architecture and saved the weights. Then processed images and text of LT-EDI Shared Task 2 along with saved models are passed through two Machine Learning

based classifiers, RandomForest, XGBoost. Between them XGBoost performed best on the test data with Macro F1 of 0.2421.

4.4 Architecture for Chinese Data

For the Chinese multimodal architecture, the data was first balanced before feature extraction. A total of 1024 features were extracted from ConvNeXt and 768 features from ChineseBERT. Each model output was passed through a linear projection layer, reducing the features to 256 per modality. These features were then concatenated to form a 512-dimensional vector and fed into a multilayer perceptron for classification. The MLP consists of two hidden layers with 128 and 32 nodes respectively, and an output layer with three nodes. This architecture achieved a macro F1 score of 0.5748 on the test data.

5 Results

The task involved two separate classification settings: one for English memes and another for Chinese memes. As shown in Table 3, different unimodal models were first evaluated to understand their individual performance. For English text, RoBERTa achieved a macro F1-score of 0.68. For Chinese text, ChineseBERT obtained a macro F1-score of 0.47. For image-based features, ConvNeXt performed strongly for both languages, reaching macro F1-scores of 0.78 for English memes and 0.75 for Chinese memes.

Table 3: Model performance comparison for English and Chinese datasets.

Model	Homophobia			Transphobia			Non_Anti_LGBT		
	P	R	F1	P	R	F1	P	R	F1
English – Image Models									
ConvNext	0.80	0.80	0.80	0.74	0.76	0.75	0.80	0.77	0.79
ResNet50	0.58	0.91	0.71	0.64	0.72	0.68	0.88	0.48	0.62
DenseNet121	0.80	0.62	0.70	0.65	0.88	0.75	0.82	0.75	0.78
InceptionV3	0.85	0.69	0.76	0.89	0.50	0.64	0.68	0.96	0.79
English – Text Models									
BERT	0.66	0.36	0.47	0.74	0.69	0.71	0.58	0.80	0.67
RoBERTa	0.58	0.54	0.56	0.77	0.79	0.78	0.70	0.72	0.71
LoRA + RoBERTa	0.60	0.50	0.55	0.80	0.55	0.65	0.59	0.80	0.68
XLM-RoBERTa	0.00	0.00	0.00	0.00	0.00	0.00	0.41	1.00	0.58
DeBERTa	0.33	0.04	0.06	0.43	0.41	0.42	0.53	0.88	0.66
English – Multimodal									
ConvNext + RoBERTa + Perceptron	0.79	0.79	0.79	0.86	0.86	0.86	0.90	0.90	0.90
ConvNext + RoBERTa + XGBoost	0.93	0.81	0.87	0.85	0.96	0.90	0.80	0.75	0.77
ConvNext + RoBERTa + RandomForest	0.92	0.75	0.83	0.79	0.96	0.87	0.86	0.75	0.80
Chinese – Image Models									
ConvNext	0.88	0.93	0.90	1.00	0.50	0.67	0.72	0.68	0.70
ResNet50	0.85	0.94	0.90	0.00	0.00	0.00	0.77	0.69	0.73
DenseNet121	0.86	0.93	0.89	0.33	0.09	0.14	0.77	0.69	0.73
Chinese – Text Models									
XLM-RoBERTa	0.74	1.00	0.85	0.00	0.00	0.00	0.00	0.00	0.00
ChineseBERT	0.76	0.97	0.85	0.75	0.33	0.46	0.25	0.03	0.05
mBERT	0.76	0.94	0.84	0.75	0.33	0.46	0.30	0.07	0.12
Chinese – Multimodal									
ConvNext + ChineseBERT + SMOTE + RandomForest	0.88	0.83	0.86	0.40	0.40	0.40	0.62	0.75	0.68

P = Precision, R = Recall, F1 = F1-score.

In English, XLM-RoBERTa showed interesting result, for both homophobia and transphobia it scored 0 in accuracy, precision and recall. it had scored perfect 1.00 in recall of Non_Anti_LGBT class. It means it had classified almost all samples as *Non_Anti_LGBT*. In Chinese, similar bias could be seen with the class *Homophobia*. There can be different possible reasons. The most probable is the fact that XLM-RoBERTa is a multi-lingual model, so it is more sensitive towards the selected hyperparameters, and size, quality of the dataset.
For the final English meme classification, several multimodal combinations were explored by combining text and image features. Among the tested approaches, the XGBoost classifier produced the best results, achieving a macro F1-score of 0.85. This model worked better than deep learning classifiers, likely because the dataset size was relatively small. For Chinese memes, the combination of ChineseBERT and ConvNeXt features along with the Random Forest classifier showed good performance. To reduce the effect of class imbalance, the SMOTE method was applied, which helped improve the detection of the minority transphobic class.
Table 4 shows the final overall result that we got on the training data. Table 5 shows our result on the test set on both Chinese and English language. For English data, it can be seen that the result on the train and test data varies a lot, whereas in Chinese the result on the test set is only slightly different than the train result. Despite the two language pipelines being very similar, there is difference between validation and test results. The possible reasons being the English dataset containing diversity of expressions, implicit sarcasm, ambiguous contexts etc. On the other hand, the Chinese dataset probably contains consistent lexical and semantic patterns, making it much easier to differentiate between each class.

Table 4: Overall performance of multimodal models across languages.

Language	Model	Accuracy	Macro Avg F1	Weighted F1
English	ConvNext + RoBERTa + XgBoost + Fine-tuning	0.86	0.85	0.86
English	ConvNext + RoBERTa + Random Forest + Fine-tuning	0.84	0.83	0.84
Chinese	ConvNext + ChineseBERT + SMOTE + Random Forest	0.79	0.65	0.80

Table 5: Evaluation on Test Data

Language	Model	Accuracy	Macro Avg F1-Score	Weighted F1-Score
English	ConvNext + RoBERTa + XGB + Fine-tuning	0.3475	0.2421	0.2791
Chinese	ConvNext + ChineseBERT + SMOTE + Random Forest	0.7500	0.5748	0.7522

6 Conclusion

In this study, we explored the problem of identifying homophobic and transphobic content in memes shared on social media. Memes are a popular way of communication, especially among younger users, but their combination of images and text makes harmful intent harder to detect. To address this challenge, we developed a multimodal approach that considers both visual and textual information in a meme. The experiments were conducted using the dataset released as part of the LT-EDI 2026 Challenge. Since the dataset is relatively small, we first pre–fine-tuned the models on the PrideMM dataset to improve their ability to understand similar content. The results show that this strategy helps the model perform better, achieving a Macro F1-score of 0.24 for English memes and 0.57 for Chinese memes. Overall, the study shows that combining multimodal learning with two-stage fine-tuning on related data can be useful for detecting harmful meme content and may help support safer online interactions.

7 Source Code

https://github.com/Saloni0000/CuriousVectors_LT-EDI-2026-B

References

Carlos Arcila-Calderón, Javier J. Amores, Patricia Sánchez-Holgado, Lazaros Vrysis, Nikolaos Vryzas, and Martín Oller Alonso. 2022. How to detect online hate towards migrants and refugees? developing and evaluating a classifier of racist and xenophobic hate speech using shallow and deep learning. *Sustainability*, 14(20).

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63.

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

Lydia Channon and Nicola Mathieson. 2025a. Automated detection of mainstreamed transphobic content on youtube. *Bulletin of Applied Transgender Studies*, 4(1-3):41–75.

Lydia Channon and Nicola Mathieson. 2025b. Automated Detection of Mainstreamed Transphobic Content on YouTube. *Bulletin of Applied Transgender Studies*, 4(1-3):41–75.

Ankita Gandhi, Param Ahir, Kinjal Adhvaryu, Pooja Shah, Ritika Lohiya, Erik Cambria, Soujanya Poria, and Amir Hussain. 2024. Hate speech detection: A comprehensive review of recent works. *Expert Systems*, 41(8):e13562.

Shrey Gupta, Pratyush Priyadarshi, and Manish Gupta. 2023. Hateful comment detection and hate target type prediction for video comments. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3923–3927.

Paulo Cezar de Q Hermida and Eulanda M dos Santos. 2023. Detecting hate speech in memes: a review. *Artificial Intelligence Review*, 56(11):12833–12851.

Saehyung Lee, Jisoo Mok, Sangha Park, Yongho Shin, Dahuin Jung, and Sungroh Yoon. 2024. Textual training for the hassle-free removal of unwanted visual data: case studies on ood and hateful image detection. *Advances in Neural Information Processing Systems*, 37:125312–125335.

Zewdie Mossie and Jenq-Haur Wang. 2020. Vulnerable community identification using hate speech detection on social media. *Information Processing Management*, 57(3):102087.

Kishore Kumar Ponnusamy, Bharathi Raja Chakravarthi, Mahesh Susaladi, Junru Ren, Prasanna Kumar Kumaresan, Premjith B, Durairaj Thenmozhi, Ruba Priyadharshini, and Subalalitha Chinnaudayar Navaneethakrishnan. 2026. Overview of Multimodal Homophobia and Transphobia Meme Classification Shared Task. In *Proceedings of the Workshop on Language Technology for Equality, Diversity, and Inclusion*. Association for Computational Linguistics.

Abir Rahali, Moulay A. Akhloufi, Anne-Marie Therien-Daniel, and Eloi Brassard-Gourdeau. 2021. Automatic misogyny detection in social media platforms using attention-based bidirectional-lstm. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2706–2711.

Hugo Rosa, Nádia Pereira, Ricardo Ribeiro, Paula Costa Ferreira, Joao Paulo Carvalho, Sofia Oliveira, Luísa Coheur, Paula Paulino, AM Veiga Simão, and Isabel Trancoso. 2019. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93:333–345.

Pradeep Kumar Roy, Snehaan Bhawal, and Chinnaudayar Navaneethakrishnan Subalalitha. 2022. Hate speech and offensive language detection in dravidian languages using deep ensemble framework. *Comput. Speech Lang.*, 75(C).

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024. Memeclip: Leveraging clip representations for multimodal meme classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17320–17332, Miami, Florida, USA. Association for Computational Linguistics.

Deepawali Sharma, Vedika Gupta, and Vivek Kumar Singh. 2023. *Detection of Homophobia amp; Transphobia in Malayalam and Tamil: Exploring Deep Learning Methods*, page 217–226. Springer Nature Switzerland.

Deepawali Sharma, Aakash Singh, and Vivek Kumar Singh. 2024. Thar-targeted hate speech against religion: A high-quality hindi-english code-mixed dataset with the application of deep learning models for automatic detection. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Elena Shushkevich and John Cardiff. 2019. Automatic misogyny detection in social media: A survey. *Computación y Sistemas*, 23(4):1159–1164.

A. Singh, V. Bansal, M. Saini, D. Sharma, and V. K. Singh. 2026. Safeplay-x: A comprehensive gameplay video dataset for violence detection with explainable deep learning applications. *Expert Systems with Applications*.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2024. Mimic: Misogyny identification in multimodal internet content in hindi-english code-mixed language. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2025. Emogif: A multimodal approach to detect emotional support in animated gifs. *IEEE Transactions on Computational Social Systems*.

Han Wang, Tan Rui Yang, Usman Naseem, and Roy Ka-Wei Lee. 2024. Multihateclip: A multilingual benchmark dataset for hateful video detection on youtube and bilibili. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7493–7502.