SAJI_English@LT-EDI 2026: Detection of Homophobia and Transphobia in Internet Memes Using Zero-Shot Learning

Jishnu Bandyopadhyay[1] **Saloni Kushwaha**[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
{jishnumsc24, salonimsc24, asingh}@cs.du.ac.in
deepawali21@bhu.ac.in

Abstract

Social media is now an important platform for communication and interaction. At the same time, the amount of abusive and harmful content online has also increased. Offensive language and hate speech are making these platforms less safe and less welcoming for users. Many of these contents include homophobic and transphobic remarks aimed at the LGBT+ community. Such behaviour damages healthy discussions and can negatively affect users. For this reason, it is important to detect these contents early so they can be flagged and removed to maintain a healthy online well-being. The issue becomes more difficult when harmful messages appear in popular formats like memes. Memes are widely used by younger users to communicate online. Because they combine images and text, detecting offensive meaning becomes challenging. In this work, we attempt to address this problem. We develop a method to identify such content using the meme dataset released for the LT-EDI 2026 challenge and secured rank 5 in the shared task. We propose a Zero-shot learning based method employing two LLMs (Qwen2.5-VL-3B-Instruct and Meta-Llama-3-8B-Instruct) to generate descriptions and classify such memes. We achieved a macro F1-score of 0.55 for the English language meme.

1 Introduction

With the rise of technology in today's world, the use of social media has increased massively. A huge portion of today's population is active on social media. It has become one of the main mediums of communication. It started as a simple break from real life, but now it is hard to live without it. Today, social media is not just a pass time for people, but also a medium of communication, source of entertainment and even a source of news for a huge majority of people (Singh et al., 2026). A number of modern jobs also revolve around social media (Singh et al., 2025a). So, it is a global platform for everyone to express themselves in any way they want, and communicate freely. But unrestricted freedom also comes with consequences. Social media is also used to target and spread sensitive and harmful content towards specific vulnerable groups of people like LGBT+ community (Yenala et al., 2018).

Although social media started as text as the medium of content shared, today it is not limited to just text, it is multimodal. Sharing all kinds of contents like images, videos, documents were made possible with time (Singh et al., 2025b). Today most of the content in social media is a combination of the previously mentioned formats. The prominent example being memes. Memes have been popularized due to their wide spectrum of ability to express. So, to identify the hateful and discriminatory contents of different types of medium, we need separate specialized tools (Fersini et al., 2022), (Singh et al., 2024).

In recent times, hateful contents in social media have skyrocketed (Weber et al., 2021). Homophobic and transphobic contents are also part of it (Chakravarthi et al., 2022a). Homophobia is a term referring to the irrational fear, discomfort or hate towards homosexual or bisexual people. Similarly, transphobia refers to fear, discomfort or hate towards transgender persons (Chakravarthi, 2024), (K et al., 2025). There are many different studies done on hate speech towards targeted individuals or vulnerable groups. And the necessity to classify hateful content in various mediums is still necessary, to make social media safe and harmless to everyone around the world. This work aims to solve the problem related to this domain using data provided by the task of LT-EDI 2026.

The paper is organised as follows : Section 2 discuss the related works in this field, Section 3 describes dataset used in the study, section 4 describes the methodlogy used in the study. The result obtained in the study are discussed in the Section 5. The learnings and findings of the work are summarised in the section 6.

2 Related Work

Many studies have been carried out to detect hate speech and offensive content in online data. With the advancement of NLP techniques, this detection process has become largely automated and more efficient (Poletto et al., 2020). Several deep learning approaches have been tested and have shown strong results in identifying hate-related content (Pamungkas et al., 2023). In addition, embedding-based and lexicon-based models have proven to be useful, especially when analyzing and identifying the specific targets of hate speech (Rawat et al.,

Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 217–221

2024). Dataset-based studies have played a key role in this research area. Early work introduced general-purpose datasets for binary classification tasks (Mathew et al., 2020). Another study presented a dataset for detecting homophobia and transphobia using YouTube data (Chakravarthi, 2024). In addition, the LT-EDI 2022 workshop offered valuable datasets along with effective methods for homophobia and transphobia detection (Chakravarthi et al., 2022b), (Sharma et al., 2022).
Several studies on hate speech detection show that simple keyword-based methods often fail to identify harmful content. These approaches rely on detecting specific offensive words, but harmful messages are not always expressed directly. In many cases, the intent is implicit, making it difficult for keyword-based systems to capture it (Fortuna and Nunes, 2018). Another limitation is that aggressive keyword filtering can produce many false positives, since some offensive terms may be used jokingly or in a mocking context rather than with harmful intent (Arango et al., 2022), (Ayo et al., 2020). Because of these challenges, recent research has explored the use of Large Language Models to classify homophobic and transphobic content (Channon and Mathieson, 2025), (Goswami et al., 2024). Building on this direction, our study bridges the gap in detecting such content by using a zero-shot learning approach that harnesses the reasoning capability of LLMs to identify harmful content even when explicit keywords are absent.

3 Dataset Description

This study uses the hateful meme dataset targeting the LGBT+ community released in the shared task of the LT-EDI 2026 Challenge (Ponnusamy et al., 2026) . The dataset contains memes that specifically target homosexual and transgender individuals. It includes data in three languages: English, Hindi, and Chinese. The English portion of the dataset is relatively balanced across classes. The Hindi data shows some level of imbalance, while the Chinese dataset has a strong class imbalance. For this work, we focused only on the English subset of the dataset. The distribution of the data is presented in Table 1.

Table 1: 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

Meme classification is a challenging task. It requires understanding both the image and the text together. Even then, the meaning can remain unclear. In multimodal settings, images and text are usually processed separately. The outputs are later combined using ensembling or feature fusion.This study follows a different approach. It uses the image reading and analysis strength of Vision Language Models (VLMs). It also uses the text understanding ability of Large Language Models (LLMs). These capabilities are combined to classify memes more effectively. The paper highlights the effectiveness of zero-shot learning with pretrained, general-purpose LLMs. The methodology is divided into two main stages. The first stage performs OCR and generates image-based analysis using a VLM. The second stage classifies the extracted text and analysis using an LLM.

4.1 Model Description

This section describes the models used in this study. For the vision–language component, the Qwen2.5-VL-3B-Instruct model developed by Alibaba Cloud is used. In the second phase, text processing and classification are performed using the Meta-Llama-3-8B-Instruct model.

4.1.1 Qwen 2.5-VL 3b Instruct

Qwen2.5-VL-3B-Instruct can handle both text and image inputs. Models of this type are known as Vision Language Models (VLMs). Qwen2.5-VL-3B-Instruct uses a separate vision encoder to process images. The encoder converts an input image into dense visual embeddings. These embeddings are then mapped to the same space as text tokens. The visual and text tokens are combined and passed to a decoder-only Transformer model. This study uses the 3-billion-parameter version due to its good performance. It is also lightweight and can run on GPUs with low VRAM.

4.1.2 Meta-Llama-3-8B-Instruct

Meta-Llama-3-8B-Instruct is a famous open-source LLM developed by Meta. The model is specifically instruction-tuned, enabling it to follow natural language prompts effectively. Meta-Llama-3-8B-Instruct is based on a decoder-only Transformer architecture, where input text is tokenized and processed through stacked self-attention layers. Its tokenizer has a vocabulary of 128k words. This LLM uses Grouped Query Attention for better inference efficiency. The original pre-training dataset is more than 15 trillion tokens long. The data is taken from publicly available sources and Llama 2. The 8 billion parameter variant is one of the most lightweight versions of Llama 3, so it can easily run on low end GPUs.

4.2 Pipeline

This section explains the pipeline of the proposed methodology. The study uses the two LLMs introduced earlier in a sequential manner. Each stage of the process is described in detail below.

4.2.1 Text Extraction and Analysis Generation

There are several ways to extract text from images. The most common method is to use OCR tools such

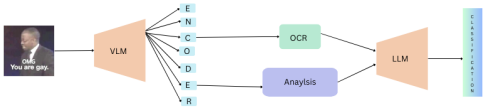

Figure 1: Proposed multimodal framework.

as Tesseract. However, this task becomes more difficult in the case of memes. The text is often spread across the image and may overlap with visual elements. This makes accurate text extraction challenging. To address this issue, this study uses a Vision Language Model, Qwen2.5-VL-3B-Instruct, for text extraction. In addition to extracting text, the model is also used to generate a semantic analysis of the meme. A zero-shot learning setup is applied for this task. No fine-tuning is performed on the VLM. Figure 1 shows the proposed multimodal framework.

To ensure transparency, the prompts used in both stages were manually designed through iterative experimentation. Different prompt variations were evaluated to determine which formulations produced the most stable and accurate outputs. During the analysis stage, different prompt structures were tested for OCR extraction, meme description generation, and contextual interpretation. The final prompt was selected because it consistently generated structured outputs containing both extracted text and semantic understanding of the meme. The final system prompt used for text extraction and analysis is presented below.

You are a meme analyzer. You should describe the meme, what is written in the meme. Finally classify the meme between HOMOPHOBIA, TRANSPHOBIA AND NON_ANTI_LGBTQ
The text in the meme
<The text in the meme >
The description of meme
<Analysis of the meme >

After extracting the text from the memes and generating the analysis, the results are stored in a CSV file. The saved data is then used as input for the next stage of the pipeline.

4.2.2 Classification

In most text classification tasks, text embeddings are used as features. These features are then passed to machine learning or deep learning models for prediction. However, simple models often fail to capture deeper meanings in text. Expressions like sarcasm and humour are particularly hard to identify, especially in data-sparse scenarios. This problem becomes more severe in data-scarce settings. To address this, this study uses the strong text understanding ability of large language models to perform the classification task.

A system prompt is designed in a structured format. It includes the extracted text and the meme analysis generated by the VLM. The LLM is instructed to classify the meme using this information. The task is limited to three possible classes. The model is asked to respond with only one word, which is the class name. No explanation or extra text is allowed. To enforce this behavior, the max_new_tokens parameter is set to 5. For consistent outputs, the temperature is set to 0. The prompt used for the classification task is shown below.

You are a strict classification model.
Task:
Classify the meme into ONLY ONE of the following classes:
HOMOPHOBIA
TRANSPHOBIA
NON_ANTI_LGBTQ
Rules:
- Output only ONE WORD, that is, one of the classes, the class name should be ALL CAPS.
- No explanation.
- No punctuation.
- No extra text.
OCR: (ocr_text)
ANALYSIS: (analysis_text)
Answer:

5 Results

The task was designed as a three-class classification problem. In the first step, we extracted the text from the memes and generated a brief analysis of their content. In the next step, the memes were classified into predefined categories. Both stages relied on a zero-shot learning approach since the available dataset was relatively small. An example of sample classification is shown in Figure 2.

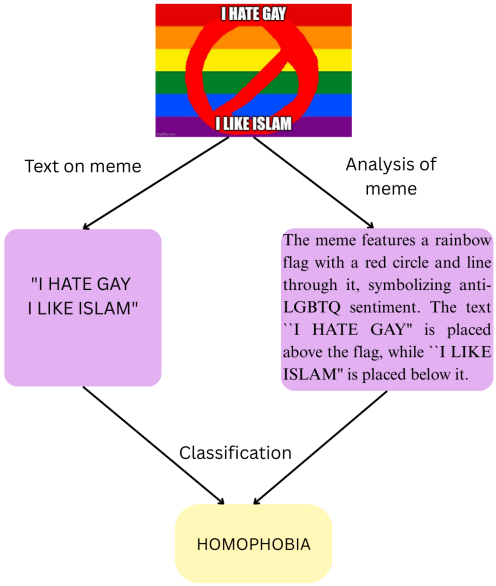

Figure 2: Example to Show Meme Analysis.

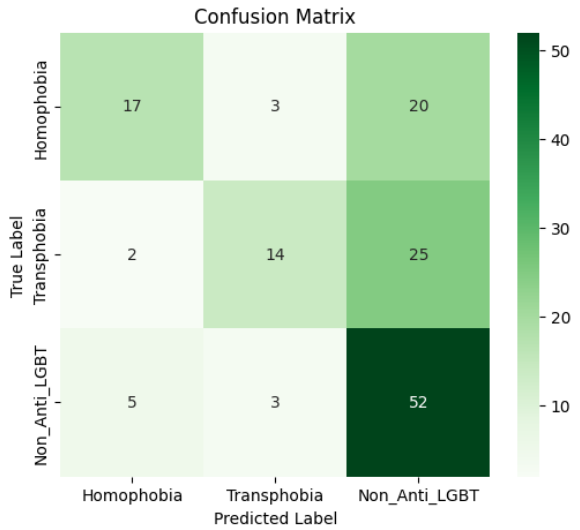

Figure 3: Confusion Matrix of Test Data.

Table 2: Evaluation on Test Data

	Precision	Recall	F1-Score	Macro F1-Score	Accuracy
Homophobia	0.71	0.42	0.53	0.55	0.59
Transphobia	0.70	0.34	0.46		
Non_Anti_LGBTQ	0.54	0.87	0.66		

Table 2 presents the detailed results obtained from the experiments, while Figure 3 shows the confusion matrix of the classification outcomes.

6 Conclusion

In this work, we addressed the problem of identifying homophobic and transphobic content in memes shared on social media. As memes combine both visual and textual information, detecting harmful intent becomes more challenging than in plain text. To address this issue, we explored a zero-shot learning-based approach that makes use of two LLMs (Qwen2.5-VL-3B-Instruct and Meta-Llama-3-8B-Instruct) to generate descriptions of the meme and then determine whether the content contains harmful intent. Our method was evaluated on the dataset released as part of the LT-EDI 2026 Challenge, where it achieved a macro F1-score of 0.55 on the English language dataset. The results show that large language models can be useful for understanding multimodal content even without task-specific training. Overall, this study highlights the potential of zero-shot approaches for moderating harmful content present on online platforms in formats such as memes and can support efforts to create safer and more respectful online spaces.

7 Source Code

https://github.com/TheRealJishnu/SAJI_English_LT-EDI-2026-B

References

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2022. Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, 105:101584.

Femi Emmanuel Ayo, Olusegun Folorunso, Friday Thomas Ibharalu, and Idowu Ademola Osinuga. 2020. Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review*, 38:100311.

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, Adeep Hande, Rahul Ponnusamy, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022a. How can we detect homophobia and transphobia? experiments in a multilingual code-mixed setting for social media governance. *International Journal of Information Management Data Insights*, 2(2):100119.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Durairaj Thenmozhi, John Philip McCrae, Paul Buitelaar, Rahul Ponnusamy, and Prasanna Kumar Kumaresan. 2022b. Overview of the shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 369–377, Dublin, Ireland. Association for Computational Linguistics.

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

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Dhiman Goswami, Sadiya Sayara Chowdhury Puspo, Md Nishat Raihan, and Al Nahian Bin Emran. 2024. MasonTigers@LT-EDI-2024: An ensemble approach towards detecting homophobia and transphobia in social media comments. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 164–172, St. Julian's, Malta. Association for Computational Linguistics.

Navya K, Hiba Sabaha, Saranya Rajiakodi, and Bhuvaneswari Sivagnanam. 2025. Detecting homophobic

and transphobic comments on social media in malayalam and english languages. *Procedia Computer Science*, 258:2479–2489. International Conference on Machine Learning and Data Engineering.

Binny Mathew, Punyajoy Saha, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2023. Towards multidomain and multilingual abusive language detection: a survey. *Personal and Ubiquitous Computing*, 27(1):17–43.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477 – 523.

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.

Anchal Rawat, Santosh Kumar, and Surender Singh Samant. 2024. Hate speech detection in social media: Techniques, recent trends, and future challenges. *WIREs Comput. Stat.*, 16(2).

Deepawali Sharma, Vedika Gupta, and Vivek Kumar Singh. 2022. Detection of homophobia & transphobia in malayalam and tamil: Exploring deep learning methods. In *International Conference on Advanced Network Technologies and Intelligent Computing*, pages 217–226. Springer.

Aakash Singh, Vinayak Bansal, Muskan Saini, Deepawali Sharma, and Vivek Kumar Singh. 2026. Safeplay-x: A comprehensive gameplay video dataset for violence detection with explainable deep learning applications. *Expert Systems with Applications*, 316:131724.

Aakash Singh, Anurag Kanaujia, and Vivek Kumar Singh. 2025a. Data to decisions: A computational framework to identify skill requirements from advertorial data. In *Advanced Network Technologies and Intelligent Computing*, pages 435–458, Cham. Springer Nature Switzerland.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2024. Mimic: Misogyny identification in multimodal internet content in hindi-english code-mixed language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2025b. Emogif: A multimodal approach to detect emotional support in animated gifs. *IEEE Transactions on Computational Social Systems*, 12:3791–3803.

Derek Weber, Mehwish Nasim, Lewis Mitchell, and Lucia Falzon. 2021. Exploring the effect of streamed social media data variations on social network analysis. *Social Network Analysis and Mining*, 11(1):62.

Harish Yenala, Ashish Jhanwar, Manoj K. Chinnakotla, and Jay Goyal. 2018. Deep learning for detecting inappropriate content in text. *International Journal of Data Science and Analytics*, 6(4):273–286.