Overview of Fourth Shared Task on Homophobia and Transphobia Span Detection in Social Media Comments

Prasanna Kumar Kumaresan[1], Bharathi Raja Chakravarthi[1], Ruba Priyadharshini[2], Paul Buitelaar[3], Malliga Subramanian[4], Kishore Kumar Ponnusamy[5]

[1]School of Computer Science, University of Galway, Ireland
[2]Gandhigram Rural Institute – Deemed to be University, Tamil Nadu, India
[3]Insight SFI Research Centre for Data Analytics,
Data Science Institute, University of Galway, Ireland
[4]Kongu Engineering College, Tamil Nadu, India
[5]Digital University of Kerala, India

Correspondence: P.Kumaresan1@universityofgalway.ie

Abstract

The rise and the intensity of harassment and hate speech on social media platforms against LGBTQ+ communities is a growing concern. This work is an initiative to address this problem by conducting a shared task focused on the detection of homophobic and transphobic content in multilingual settings. The task comprises two subtasks: (1) multi-class classification of content into homophobia, transphobia, or non-anti-LGBT+ categories across eight languages and (2) span-level detection to identify specific toxic segments within comments in English, Tamil, and Marathi. This initiative helps the development of explainable and socially responsible AI tools for combating identity-based harm in digital spaces. Multiple teams registered for the task; however, only two teams submitted their results, and the results were evaluated using the macro F1 score.

1 Introduction

Homophobia and transphobia refer to harmful attitudes and prejudices directed toward individuals who identify as homosexual or transgender[1] (Hill, 2003; O'Donohue and Caselles, 1993; Nagoshi et al., 2008). While the terms may linguistically suggest irrational fear, they more accurately encompass a spectrum of negative biases and discriminatory behaviors against people who are lesbian, gay, bisexual, or transgender (Rollè et al., 2014). These biases can manifest in various forms, ranging from subtle expressions such as derogatory language to overt acts of hostility and aggression, contributing significantly to the marginalization and emotional distress experienced by LGBTQ+ individuals (Moagi et al., 2021).

Recently, the growth of social media has both amplified these challenges and created new avenues for their expression (Fuchs, 2014; Chakravarthi et al., 2022). While these platforms foster connection and community building, they have also become grounds for the spread of toxic language, including hate speech targeting LGBTQ+ communities (Kumaresan et al., 2023; Calderón et al., 2024). According to the European Union, 50% of LGBT persons have been victims of hate speech or hate crime [2]. Such homophobic and transphobic content online not only reinforces societal prejudices but also inflicts psychological harm (Newcomb and Mustanski, 2010). Therefore, the ability to detect and address such harmful language in social media content is essential for cultivating safer, more inclusive digital environments (Chakravarthi, 2024).

This shared task addresses the problem of homophobia and transphobia detection in social media comments. It aims to promote research into the automatic identification and classification of homophobic and transphobic language, with a particular focus on multilingual and under-resourced language contexts. The shared task comprises two components: comment-level classification (Kumaresan et al., 2023) and span-level detection (Kumaresan et al., 2025). This involves highlighting the exact phrases that serve as evidence for the classification, enabling a more fine-grained and interpretable analysis. Span detection is particularly valuable for building explainable NLP systems that not only flag harmful content but also provide trans-

[1] https://reportandsupport.qmul.ac.uk/support/what-is-homophobia-transphobia-acephobia-and-biphobia

[2] https://fra.europa.eu/sites/default/files/fra_uploads/1226-Factsheet-homophobia-hate-speech-crime_EN.pdf

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 228–233
September 9, 2025 ©2025 Association for Computational Linguistics

parent justifications for their decisions (Naim et al., 2022).

The dataset used for this task is derived from the manually annotated homophobia/transphobia content, which includes YouTube comments labeled at the comment level. Participants are encouraged to develop robust NLP systems capable of accurately identifying and categorizing hate speech targeting LGBTQ+ individuals. By tackling both classification and span detection, this shared task provides a platform for the NLP community to advance techniques for harmful content detection while fostering socially responsible NLP research across diverse linguistic and cultural settings.

In the upcoming section, we will describe the task description, dataset statistics, and participants' methodology towards the investigation of homophobia and transphobia detection from the YouTube comments on Dravidian languages.

2 Related Works

Span detection, also known as span-based classification or span identification, involves pinpointing the specific segments of a text that contain harmful or toxic content, rather than labeling the entire text as toxic (Pavlopoulos et al., 2021). This approach is particularly valuable in scenarios where only a small portion of a comment or post contains offensive language, while the remainder may be benign or contextually neutral (Gu et al., 2022). Traditional text classification models typically assign a single label to the entire input, which can be limiting in practical content moderation settings. Flagging an entire message as toxic based on a minor fragment may lead to unnecessary censorship and hinder constructive discourse.

Recent research in hate speech detection has increasingly emphasized the importance of explainability and precision (Sawant et al., 2024; Calabrese et al., 2024). Span-level annotations offer moderators actionable insights by highlighting the exact portions of the text that violate community guidelines, thereby streamlining the moderation process and enabling more targeted interventions (Mathew et al., 2021). This is especially crucial in social media contexts where high volumes of user-generated content make manual review inefficient.

In the context of homophobia and transphobia, span detection plays a critical role in identifying instances of identity-based harm (Zhou et al., 2023). Recent studies such as (Kumaresan et al., 2024) have explored the use of fine-grained annotations to detect hate speech against LGBTQ+ individuals, highlighting the need for datasets and models that capture identity-specific slurs and implicit hate spans. Studies (Condom Tibau et al., 2025; Chakravarthi et al., 2024) further illustrate the challenges in reliably detecting toxic content targeted at LGBT communities, showing that span-based approaches can improve both precision and fairness in these cases. These advances underscore the value of targeted span detection for moderating homophobic and transphobic content, offering more transparent and inclusive systems for content moderation.

Languages	Set	H	T	N
English	Train	179	7	2,978
	Dev	42	2	748
	Test	55	4	931
Tamil	Train	453	145	2,064
	Dev	118	41	507
	Test	152	47	634
Malayalam	Train	476	170	2,468
	Dev	197	79	937
	Test	140	52	674
Telugu	Train	2,907	2,647	3,496
	Dev	588	605	747
	Test	624	571	744
Kannada	Train	2,765	2,835	4,463
	Dev	585	617	955
	Test	599	606	951
Gujarati	Train	2,267	2,004	3,848
	Dev	498	454	788
	Test	510	436	794
Hindi	Train	45	92	2,423
	Dev	2	13	305
	Test	3	10	308
Marathi	Train	551	377	2,572
	Dev	129	80	541
	Test	112	69	569

Table 1: Multilingual classification (Task 1) dataset statistics (H-Homophobia, T-Transphobia, and N-Non-anti-LGBT+ content)

3 Task Description

We organized the shared task on homophobia & transphobia with around two subtasks.

- *Subtask 1*: Homophobia & Transphobia Multilingual Classification Task

  - Objective: Classify comments into three categories: Homophobia, Transphobia, and None of the Above.
  - Languages: This task will be conducted in multiple languages, specifically English, Tamil, Malayalam, Hindi, Gujarati, Telugu, Kannada, and Marathi.
  - Special Focus on Tulu: Given the scarcity of resources like annotated corpora for under-resourced languages like Tulu, this task presents a unique challenge. We have introduced a code-mixed Tulu dataset specifically designed to detect homophobic and transphobic content. This dataset aims to promote research in few-shot learning, pushing the boundaries of what's possible in language processing for low-resource contexts.

- *Subtask 2*: Homophobia & Transphobia Span Detection
  - Objective: Identify specific spans within comments that contain instances of homophobia and transphobia.
  - Languages: English, Tamil, and Marathi.
  - Details: Participants will be provided with comments and are required to classify these comments at the span level. This task requires a deeper level of text understanding and precision, as participants must discern and highlight the textual evidence for homophobia or transphobia within the comments.

Overall, these tasks are designed not only to address significant technical challenges in the field of NLP but also to contribute to social good by identifying and mitigating harmful content directed at the LGBTQ+ community in diverse linguistic contexts.

4 Dataset Statistics

Social media platforms Twitter, Facebook, and YouTube use user-generated content to shape public opinion, which affects how people perceive things and how they view others. Recognizing the growing need for automated tools to extract emotions and detect harmful or irrelevant content online, particularly on platforms like YouTube, where user comments are rapidly increasing, we focused on content relevant to the LGBTQ+ community, who frequently engage with such platforms to express their views on various topics.

Languages	**Set**	**H**	**T**	**N**
Tamil	Train	188	75	137
	Test	73	36	63
English	Train	117	39	44
	Test	49	17	20
Marathi	Train	253	119	123
	Test	108	53	52

Table 2: Span Detection (Task 2) dataset statistics (H-Homophobia, T-Transphobia, and N-None of above)

We gathered a multilingual collection of YouTube comments about LGBTQ+ for Task 1. We protected individual privacy by not including personal stories from LGBTQ+ individuals in our collection. Using the YouTube Comment Scraper tool, we collected comments and manually annotated them with one of three labels: homophobic, transphobic, and non-anti-LGBT+ content. The final dataset languages - English, Tamil, Malayalam, Telugu, Kannada, Gujarati, Hindi, and Marathi were annotated following the guidelines outlined in the dataset paper (Kumaresan et al., 2023). The distribution of annotated labels across all languages appears in Table 1.

For Task 2, we extended our efforts by annotating spans of text within comments that explicitly or implicitly expressed homophobia or transphobia (Kumaresan et al., 2025). These span-level annotations were carried out in three languages, Tamil, English, and Marathi, using the sequence labeling approach implemented in the open-source annotation tool Doccano. We focused on marking only those portions of text that conveyed discriminatory attitudes, allowing us to take a targeted and strategic annotation approach. Table 2 shows the dataset statistics for span annotations across the three categories: Homophobia (H), Transphobia (T), and Non-anti-LGBT+ content (N).

5 Participants Methodology

We organized a shared task focused on addressing harmful content that targets LGBTQ+ individuals through two essential subtasks. The participants used multiple machine learning and deep learning approaches to tackle these subtasks, especially when working with low-resource and multilingual data.

The *SKV TRIO team* (Vignesh et al., 2025) used a combination of BERT and TF-IDF embeddings for Task 1. The team used BERT and TF-IDF embeddings for each input before applying dimensionality reduction to TF-IDF embeddings to match BERT's dimensions. The system combined these embeddings to create a single feature representation, which served as input for training a random forest classifier. The method united semantic depth with statistical feature patterns to produce an interpretable and efficient computational solution.

The *KEC-Elite-Analysts* team used multiple deep learning models to solve task 1 by classifying homophobia and transphobia. The architecture used bidirectional LSTM and GRU models to extract sequential and contextual language patterns and class weights to handle class imbalance. A TextCNN module to detect local n-gram features indicative of toxic expressions. A multilayer perceptron (MLP) trained on averaged word embeddings to incorporate semantic information into the final prediction.

The models were designed to generalize across multiple languages, including low-resource and code-mixed languages such as Tamil. This multilingual focus ensured robust performance in linguistically diverse and underrepresented languages.

6 Result and Discussion

A total of 30 participants registered for our shared task. Nevertheless, only two teams submitted results for Task 1, and no submissions were received for Task 2. The fact that Task 2 required identifying specific spans of homophobic and transphobic content may have contributed to its lack of submissions. This probably required more domain knowledge and work, which might have made it difficult for participants to finish in the allotted time.

The outcomes of *Task 1* are displayed. The macro F1 scores for the two participating teams, SKV TRIO and KEC-Elite-Analysts, across the supported languages are shown in Table 3. To take into consideration the dataset's multilingual nature and multi label task, the evaluation was carried out independently for each language with macro F1 score. Because it computes the F1 score for each class separately and then averages them, treating all classes equally regardless of size, we decided to use the macro F1 score to assess the ranklist result.

In the majority of languages, including Gujarati (0.86), Telugu (0.87), and Kannada (0.81), the *SKV TRIO* team received the highest scores. Their method of training a random forest classifier by combining BERT and TF-IDF embeddings seems to have successfully identified both statistical and semantic patterns in the data. Low-resource and morphologically rich languages may have benefited most from this hybrid embedding approach, as term-level distinctions unique to hate speech patterns are reinforced by TF-IDF, while pre-trained contextual models such as BERT can offer general language understanding. The model's strong performance in languages with limited resources and varying the dimensionality alignment between embeddings, which also helped the model generalize better across a variety of linguistic structures.

The *KEC-Elite-Analysts* team outperformed the SKV TRIO team in English (0.40) and Tamil (0.74), demonstrating notable competence in those languages. Their system used an MLP trained on averaged word embeddings in conjunction with a deep learning ensemble comprising Bidirectional LSTM, GRU, and TextCNN components. This architecture works well with languages like English, where pre-trained embeddings and deep learning models typically perform reliably due to an abundance of resources, and Tamil, where code-mixing and sequential dependencies are common. Their system was able to capture subtle patterns in sentence structure, particularly in high-resource or semi-structured languages, because of the ensemble design and the use of class weights to address label imbalance.

These findings show that various modeling approaches obtained performance differences across languages, which may have been caused by the variety of languages and the accessibility of data. While KEC-Elite-Analysts' deep learning ensemble approach proved successful in identifying patterns in more resource-intensive or frequently used languages like English and Tamil, SKV TRIO's fusion-based feature engineering demonstrated superior generalization across a wider range of languages. The findings highlight how crucial model diversity and adaptability are, especially when working in environments with limited resources and code-mixed languages. They also highlight the need for more research into span-level detection, since future versions of the task might benefit from longer timeframes, more annotation support, or easier baselines for span identification to reduce the barrier to entry.

Team Name	**Task 1: Languages - Macro F1 Score**							
	English	**Gujarathi**	**Hindi**	**Tamil**	**Telugu**	**Marathi**	**Malayalam**	**Kannada**
SKV TRIO (Vignesh et al., 2025)	0.34	**0.86**	**0.33**	0.37	**0.87**	0.29	**0.40**	**0.81**
KEC-Elite-Analysts (Run 1)	**0.40**	-	-	**0.74**	-	**0.52**	-	-

Table 3: Results from Task 1 showing macro F1 scores by language for each participating team (bold values indicate the highest score per language).

7 Conclusion

In this shared task, we addressed the challenge of two sub-tasks, which are detecting homophobia and transphobia classification and span identification through multilingual and low-resource languages. A total of 30 participants were registered, only two teams submitted the results for Task 1, and no submissions were received for Task 2, likely due to the complexity of span annotation and the need for domain-specific understanding within a limited timeframe. The classification results showed the efficacy of various modeling approaches, with deep learning ensembles performing well in high-resource languages and hybrid embedding approaches excelling in low-resource contexts. These results emphasize how crucial flexible, language-sensitive models are for identifying harmful content. Although span detection remains a challenging and underexplored area, specifically in low-resource, it is critical for the development of explainable and culturally aware moderation systems. Future iterations of this task should aim to reduce entry barriers and further promote research in inclusive and socially responsible NLP.

Acknowledgments

Author Bharathi Raja Chakravarthi was funded by a research grant from Research Ireland under grant number SFI/12/RC/2289_P2 (Insight). This work is also supported by the Centre for Research Training in Artificial Intelligence grant number 18/CRT/6223, and a grant from the College of Science and Engineering, University of Galway, Ireland.

References

Agostina Calabrese, Leonardo Neves, Neil Shah, Maarten Bos, Björn Ross, Mirella Lapata, and Francesco Barbieri. 2024. Explainability and hate speech: Structured explanations make social media moderators faster. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 398–408, Bangkok, Thailand. Association for Computational Linguistics.

Carlos Arcila Calderón, Patricia Sánchez Holgado, Jesús Gómez, Marcos Barbosa, Haodong Qi, Alberto Matilla, Pilar Amado, Alejandro Guzmán, Daniel López-Matías, and Tomás Fernández-Villazala. 2024. From online hate speech to offline hate crime: the role of inflammatory language in forecasting violence against migrant and lgbt communities. *Humanities and Social Sciences Communications*, 11(1):1369. Published: October 15, 2024.

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

Bharathi Raja Chakravarthi, Prasanna Kumaresan, Ruba Priyadharshini, Paul Buitelaar, Asha Hegde, Hosahalli Shashirekha, Saranya Rajiakodi, Miguel Ángel García, Salud María Jiménez-Zafra, José García-Díaz, Rafael Valencia-García, Kishore Ponnusamy, Poorvi Shetty, and Daniel García-Baena. 2024. Overview of third shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 124–132, St. Julian's, Malta. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 2022. 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.

Jordi Guillem Condom Tibau, Angelina Voggenreiter, elena pavan, and Jürgen Pfeffer. 2025. Prevalence, substance and responses to hate speech against lgbtq communities on tiktok. *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):430–442.

Christian Fuchs. 2014. *Social Media: A Critical Introduction*. SAGE Publications Ltd, London.

Weiwei Gu, Boyuan Zheng, Yunmo Chen, Tongfei Chen, and Benjamin Van Durme. 2022. An empirical study on finding spans. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3976–3983, Abu

Dhabi, United Arab Emirates. Association for Computational Linguistics.

Darryl B Hill. 2003. Genderism, transphobia, and gender bashing: A framework for interpreting anti-transgender violence. In *Understanding and dealing with violence: A multicultural approach*, pages 113–136. SAGE Publications, Inc.

Prasanna Kumar Kumaresan, Devendra Deepak Kayande, Ruba Priyadharshini, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2025. Homophobia and transphobia span identification in low-resource languages. *Natural Language Processing Journal*, 12:100169.

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Ruba Priyadharshini, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2023. Homophobia and transphobia detection for low-resourced languages in social media comments. *Natural Language Processing Journal*, 5:100041.

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Dhruv Sharma, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. Dataset for identification of homophobia and transphobia for Telugu, Kannada, and Gujarati. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4404–4411, Torino, Italia. ELRA and ICCL.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

M.M. Moagi, A.E. van Der Wath, P.M. Jiyane, and R.S. Rikhotso. 2021. Mental health challenges of lesbian, gay, bisexual and transgender people: An integrated literature review. *Health SA Gesondheid*, 26:1487.

Julie L Nagoshi, Katherine A Adams, Heather K Terrell, Eric D Hill, Stephanie Brzuzy, and Craig T Nagoshi. 2008. Gender differences in correlates of homophobia and transphobia. *Sex roles*, 59:521–531.

Jannatun Naim, Tashin Hossain, Fareen Tasneem, Abu Nowshed Chy, and Masaki Aono. 2022. Leveraging fusion of sequence tagging models for toxic spans detection. *Neurocomputing*, 500:688–702.

Michael E. Newcomb and Brian Mustanski. 2010. Internalized homophobia and internalizing mental health problems: A meta-analytic review. *Clinical Psychology Review*, 30(8):1019–1029.

William O'Donohue and Christine E Caselles. 1993. Homophobia: Conceptual, definitional, and value issues. *Journal of Psychopathology and Behavioral Assessment*, 15:177–195.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.

Luca Rollè, Piera Brustia, and Angela Caldarera. 2014. *Homophobia and Transphobia*, pages 2905–2910. Springer Netherlands, Dordrecht.

Madhuri Sawant, Arjumand Younus, Simon Caton, and Muhammad Atif Qureshi. 2024. Using explainable ai (xai) for identification of subjectivity in hate speech annotations for low-resource languages. In *Proceedings of the 4th International Workshop on Open Challenges in Online Social Networks*, OASIS '24, page 10–17, New York, NY, USA. Association for Computing Machinery.

Konkimalla Laxmi Vignesh, Mahankali Sri Ram Krishna, Dondluru Keerthana, and Premjith B. 2025. Skv trio@lt-edi-2025: Hybrid tf-idf and bert embeddings for multilingual homophobia and transphobia detection in social media comments. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Linda Zhou, Andrew Caines, Ildiko Pete, and Alice Hutchings. 2023. Automated hate speech detection and span extraction in underground hacking and extremist forums. *Natural Language Engineering*, 29(5):1247–1274.