Findings of Shared Task on Counter Narrative Generation on Homophobic and Transphobic Comments

Prasanna Kumar Kumaresan[1], Praveen Prasannan[1], Tanay Singh[1], Ruba Priyadharshini[2], Subalalitha Chinnaudayar Navaneethakrishnan[3], Saranya Rajiakodi[4], Paul Buitelaar[5], Bharathi Raja Chakravarthi[1]

[1]Data Science Institute, University of Galway, Ireland
[2]Gandhigram Rural Institute – Deemed to be University, Tamil Nadu, India
[3]SRM Institute of Science and Technology, Tamil Nadu, India
[4]Central University of Tamil Nadu (CUTN), Tamil Nadu, India
[5]Data Science Institute, University of Galway, Ireland

Correspondence: P.Kumaresan1@universityofgalway.ie

Abstract

Online platforms continue to witness harmful expressions targeting LGBTQ+ individuals, particularly in the form of homophobic and transphobic comments. While detection of such content has received substantial attention, generating constructive counter-narratives remains comparatively underexplored. In this shared task, we focus on counter-narrative generation in English and Tamil. Participants were provided with social media comments labeled as homophobic or transphobic and were required to generate respectful, contextually appropriate responses that challenge prejudice and promote empathy. Systems were evaluated using both reference-based metrics (Distinct-2 and BERTScore-F1) and rubric-based human evaluation metrics measuring politeness (PRS), quality (QS), and contextual coherence (CCNC). The results demonstrate variation in system performance across languages, with English systems showing stronger lexical diversity and Tamil systems excelling in politeness and contextual coherence. This paper presents dataset statistics, evaluation methodology, system performance analysis, and key observations from the shared task.

Keywords: Counter-narratives, Hate speech detection, Homophobia and transphobia, Span detection, Multilingual NLP, Large language models

1 Introduction

The Internet serves as one of the primary means of communication, socialization, and community-building for many individuals worldwide. The Internet's role in propagating hostile speech towards marginalized communities is a major issue (Berger et al., 2022; Keighley, 2022; De Ridder and Van Bauwel, 2015). Specifically, homophobic and transphobic bullying towards sexual and gender diversity individuals remains a prevalent issue across various social media platforms, where such speech is frequently exhibited using terms such as slurs, stereotypes, and mockery, which in turn lead to the marginalization and stigmatization of individuals with sexual and gender diversity identities (Hill, 2002; Nagoshi et al., 2008; O'Donohue and Caselles, 1993). As a result, the field of natural language processing (NLP) has been focusing on the identification of hate speech and abusive language, with numerous recent studies operating within it (Sai and Sharma, 2021; Gao et al., 2020; Díaz-Torres et al., 2020). The initial studies were grounded on the conventional machine learning approaches, which make use of lexical and syntactic features, while the more recent studies make use of transformer-based approaches such as BERT and multilingual models, which lead to a significant improvement in the accuracy of the resultant models for hate speech identification on various multilingual and multi-domain datasets. While the identification of hate speech is considered an important step in the direction of resolving the issue of online bullying, it is solely insufficient in tackling the more complex social issues that are strongly associated with the proliferation of harmful language online, such as the need for comprehensive educational programs, community engagement, and policy changes that address the root causes of online harassment (Norton, 1997; Schope and Eliason, 2004).

Counter-narratives have in recent years been proposed as a way to combat hate speech on the web. A counter-narrative is a message that is intended to counteract or neutralize the content. Here, counter-narratives are meant to be respectful while

Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 118–127

at the same time challenging prejudice and trying to elicit a more empathetic attitude and self-reflection in users (Benesch, 2014; Garland et al., 2020; Fanton et al., 2021). The effect of constructive counter-speech was proven to significantly reduce cyberbullying and promote online discussions in a more positive direction. The challenge with good counter-narratives is that they have to be relevant to the original comment and must not contain rude or threatening language (Doğanç and Markov, 2023). While there is a significant body of work addressing the issue of counter-narratives in the context of English, the linguistic resources available for other languages are typically limited, and for low-resource languages such as Tamil, there is generally a lack of well-annotated resources that would facilitate the creation of an effective multilingual counter-narrative generation system (Tekiroğlu et al., 2020).

To promote research, we organize a shared task on counter-narrative generation for homophobic and transphobic comments. This shared task consists of two subtasks. Span detection identifies the word or phrase in the original input that may contain homophobic or transphobic terms, while counter-narrative generation produces a narrative responding to a given claim while refraining from using derogatory terms. These tasks are performed on a diverse set of languages, including English, Tamil, and Hindi, for the span detection task and English and Tamil for the counter-narrative generation task. The shared task focused on hate speech, counter-speech, and multilingual language processing. This task aimed to provide a testbed for researching models that can identify and locate hate speech, as well as generate counter speech to help mitigate its negative effects by offering relevant training data with annotations, evaluation metrics, and an overview of the developed systems and their outputs. This paper presents an overview of the dataset, the systems developed for this task, the evaluation methods, and the key findings.

2 Related Work

2.1 Hate Speech Detection and Span Identification:

Automated hate speech detection has been widely studied across languages, with early approaches relying on traditional machine learning models and lexical features. More recent work makes use of transformer-based architectures such as BERT and XLM-R for improved contextual modeling. A major step toward explainable hate speech detection was introduced by HateXplain (Mathew et al., 2021), highlighting the spans responsible for hateful content. This span-level formulation helps not only to classify content but also to identify the precise textual boundaries of harmful expressions. Such datasets reduce over-prediction, directly motivating span-detection subtasks in shared tasks. Multilingual hate speech detection has gained increasing importance , especially for low-resource and code-mixed languages. Ranasinghe and Zampieri (2020) demonstrated the effectiveness of transformer-based multilingual models for cross-lingual hate speech detection. Similarly, several LT-EDI shared tasks (Chakravarthi, 2024; Chakravarthi et al., 2024, 2023, 2022) have advanced research in Tamil, Hindi, and other Indian languages, highlighting challenges such as class imbalance, dialectal variation, and limited annotated resources. These works establish the foundation for Subtask 1, where span-level detection in English, Tamil, and Hindi requires both contextual modeling and cross-lingual robustness (Kumaresan et al., 2025a).

2.2 Counter Speech and Counter-Narrative Generation

While detection has received sufficient attention, mitigation through counter-narrative generation remains comparatively underexplored. Counter speech aims to address harmful content through constructive, empathetic, and non-aggressive responses. Chung et al. (2021) explored automatic counter-narrative generation using transformer-based models trained on curated counter-speech datasets. Fanton et al. (2021) proposed human-in-the-loop generation strategies, combining expert-crafted responses with neural generation to ensure politeness and factual grounding. These studies emphasizes the importance of tone, empathy, and contextual appropriateness in counter speech. Tekiroğlu et al. (2020) extended this direction to multilingual settings, proving that counter-narrative systems must balance cultural sensitivity with linguistic variation. Recent research suggests that polite and thoughtful responses are more effective in lowering hostility than confrontational reactions. These works directly motivate Subtask 2, where systems are required to generate respectful and contextually coherent counter-narratives in English and Tamil.

Table 1: Dataset Statistics for Counter-Narrative Generation on Homophobic and Transphobic Comments Shared Tasks

Task	Language	Split	Homophobia	Transphobia	None of the above	Total
Subtask 1	Tamil	Train	188	75	137	400
		Test	73	36	–	109
	English	Train	117	39	44	200
		Test	49	17	–	66
	Hindi	Train	20	34	–	54
		Test	03	10	–	13
Subtask 2	Tamil	Train	342	458	–	800
		Test	73	36	–	109
	English	Train	1,044	756	–	1,800
		Test	49	17	–	66

2.3 Evaluation of Generated Counter-Narratives

Evaluating generated counter-narratives has unique challenges, as lexical overlap alone cannot capture empathy or appropriateness. BERTScore (Zhang et al., 2019) introduced a contextual embedding-based metric that measures semantic similarity between generated text and references. It has become a standard metric for generation tasks where paraphrasing is common. Diversity-based metrics such as Distinct-N (Li et al., 2016) measure lexical variation and help prevent repetitive or templated outputs. Recent work (Chiang and Lee, 2023) examined the reliability of LLM-based evaluation frameworks, showing that rubric-based scoring can approximate human judgments in open-ended generation tasks. Such approaches enable scalable assessment of politeness, coherence, and quality attributes critical to counter speech. The combination of reference-based metrics Distinct-2, BERTScore-F1 and rubric-based scoring aligns with emerging best practices in evaluating socially sensitive generative systems (Prasannan et al., 2025).

2.4 Multilingual and Low-Resource Generation

Low-resource languages such as Tamil pose additional challenges for both detection and generation. Cross-lingual models like XLM-R (Conneau et al., 2020) provide shared multilingual representations that enable transfer learning across languages. Large-scale multilingual systems such as No Language Left Behind (Costa-Jussà et al., 2022) demonstrate that scaling multilingual pretraining improves performance in low-resource languages. Multilingual instruction tuning (Zhang et al., 2023) has shown that LLMs can generalize better for low-resource languages when exposed to diverse task instructions during training. These advancements support the multilingual design of the shared task and explain performance variations observed between English and Tamil systems.

3 Task Description

The shared task on counter-narrative generation on homophobic and transphobic comments addresses the identification and mitigation of hate speech targeting LGBTQ+ communities. Homophobia and transphobia represent harmful forms of online discourse that marginalize individuals based on sexual orientation and gender identity. To promote research in both detection and response generation, the shared task is organized into two subtasks[1].

Subtask 1: Homophobia and Transphobia Span Detection. This subtask focuses on fine-grained identification of harmful content. Given a social media comment, systems are required to detect and extract the exact textual spans that express homophobic or transphobic content. Unlike sentence-level classification, this formulation emphasizes precise boundary detection, encouraging systems to minimize over-prediction while accurately capturing abusive expressions. The task is conducted in three languages: English, Tamil, and Hindi. Systems are evaluated using standard classification metrics, including accuracy, macro-precision, macro-recall, macro-F1 (submissions were ranked based on), weighted-precision,

[1] https://sites.google.com/view/lt-edi-2026/home

Table 2: Overview of participating systems for Task 1: Homophobia and Transphobia Span Detection

Team Name	Base Model / LLM	Technique				
		Seq Label FT	Prompt Eng.	Few-shot	Post-processing	Structured Output
DuoNova	Transformer-based LM	✓			✓	
TeamV	Qwen3-Max		✓	✓	✓	✓

Table 3: Overview of participating systems for Task 2: Counter-Narrative Generation

Team Name	Base Model / LLM	Technique					
		Seq2Seq FT	Prompt Eng.	Few-shot	Instruction Tuning	Prompt Optimization	Decoding Strategy
DLRG	TF-IDF + Classical ML	✓					
Amritha	Llama 3.2 / Gemini		✓	✓	✓		
NEUNI	DSPy + LLM		✓		✓	✓	
JusticeBots	ChatGPT		✓		✓		
TeamV	Qwen3-Max		✓	✓	✓		
SigJBS	Gemma 3 (QLoRA)	✓	✓		✓		✓
RespectNLP	Seq2Seq Transformer	✓	✓				✓
DuoNova	FLAN-T5	✓					✓

weighted-recall, and weighted-F1.

Subtask 2: Counter-Narrative Generation. This subtask moves beyond detection toward mitigation. Given a comment containing homophobic or transphobic content, systems must generate a constructive counter-narrative that challenges harmful claims while remaining respectful, empathetic, and contextually coherent. The generated response should avoid hostility and instead promote inclusivity and meaningful dialogue. This subtask is conducted in English and Tamil. Evaluation includes both:

1. **Reference-Based Metrics**
    - **Distinct-2**: Measures bigram-level diversity (higher indicates less repetition).
    - **BERTScore-F1**: Measures semantic similarity between system outputs and reference counter-narratives.
2. **Rubric-Based Metrics**: Evaluated using LLM-captured properties that automatically overlap-based metrics on a 0–2 scale.
    - **PRS**: Politeness and Respectful Score.
    - **QS**: Quality Score.
    - **CCNC**: Contextual Counter-Narrative Coherence Score.

For ranking purposes, all evaluation scores were converted into percentages and averaged across the considered metrics for each team. The final leaderboard was determined based on this overall average score.

4 Dataset

The dataset for the shared task consists of multilingual social media comments annotated for homophobia and transphobia. The detailed distribution of instances across languages, splits, and categories is presented in Table 1. The dataset is designed to support both fine-grained span detection (Subtask 1) and counter-narrative generation (Subtask 2), with language coverage varying across subtasks.

For Subtask 1 (Span Detection), the dataset includes comments annotated at the span level under three categories: Homophobia, Transphobia, and None of the above. The data is available in Tamil, English, and Hindi. In Tamil, the training set contains 188 homophobia instances, 75 transphobia instances, and 137 none-of-the-above instances, totaling 400 comments; the test set includes 73 homophobia and 36 transphobia instances (109 total). In English, the training set consists of 117 homophobia, 39 transphobia, and 44 none-of-the-above instances (200 total), while the test set includes 49 homophobia and 17 transphobia instances (66 total). In Hindi, the training data contains 20 homophobia and 34 transphobia instances (54 total), and the test set includes 3 homophobia and 10 transphobia instances (13 total). The distribution reflects natural class imbalance and linguistic variation across languages (Kumaresan et al., 2025b).

For Subtask 2 (Counter-Narrative Generation), the dataset includes only comments containing homophobic or transphobic content, as the task focuses on generating constructive counter-speech. In Tamil, the training set contains 342 homophobia and 458 transphobia instances (800 total), with 73 homophobia and 36 transphobia instances in

the test set (109 total). In English, the training set consists of 1,044 homophobia and 756 transphobia instances (1,800 total), and the test set includes 49 homophobia and 17 transphobia instances (66 total). None-of-the-above category is excluded in this subtask, ensuring that all inputs require the generation of a counter-narrative response.

5 Participants Methodology

For Subtask 1, teams used a variety of approaches to extract the spans of homophobic and transphobic content. Tables 2 and 3 provide a detailed comparison of techniques. The **DuoNova** team (S et al., 2026), the subtask was addressed as a token-level sequence labeling task using a transformer. The comments were tokenized and the pre-trained language model was fine-tuned to classify each token as hateful or non-hateful. The character-level annotations were mapped to the corresponding token-level labels in the model to carry out the training. The fine-tuned model was trained on cross-entropy loss with the AdamW optimizer. During inference, all the tokens predicted as hateful were mapped back to their corresponding character locations. This resulted in exact locations and highlighting of the homophobic and transphobic content within the comments.

Team V (Ulli and Kumari, 2026) used a large language model and applied the technique of in-context learning using the instruction-tuned Qwen3-Max model [2]. The prompt was a 10-shot balanced set of English and Hindi (with romanization). The model was asked to output the minimal hateful span and to categorize it, with the expected output being in the form of a JSON format. The required character-level spans were obtained by using a multi-stage post-processing pipeline consisting of exact string search, whitespace-preserving string search, case-insensitive string search and fuzzy substring search using SequenceMatcher.

For subtask 2, the teams experimented with various techniques to produce constructive counter-narratives to answer homophobic and transphobic remarks. All the teams utilized large language models or the transformer-based sequence-to-sequence architecture to design appropriate counter-narratives. For this task, **DuoNova** (S et al., 2026) formulated the problem as a supervised sequence-to-sequence generation problem by utilizing a transformer-based encoder–decoder architecture. They employed the FLAN-T5 model (Chung et al., 2024), which was then fine-tuned with a set of paired training examples, which included the original hate speech comments and corresponding counter-narratives. The input texts were cleaned and tokenized by using the subword tokenizer from the model. The training was carried out by utilizing teacher forcing with cross-entropy loss and the optimizer was AdamW. The fine-tuned model was employed during the inference phase to produce appropriate counter-narratives by using greedy decoding or beam search.

NEUNI (Gajawada et al., 2026) proposed a prompt optimization strategy based on the DSPy MIPRO optimizer. The method applied Bayesian optimization to explore the prompt space and identify instructions that maximize evaluation criteria. Candidate prompts were evaluated using an LLM-based rubric-aligned evaluator, and the prompt achieving the highest validation performance was used for generating counter-narratives on the test data. **RespectNLP** (Priya and Bharathi, 2026) utilized a pretrained multilingual sequence-to-sequence transformer model combined with instruction-based prompting. The prompts explicitly guided the model to produce respectful, empathetic, and context-aware responses aligned with PRS, CCNC, and QS criteria. Each comment was converted into a structured prompt emphasizing courteous disagreement and constructive dialogue. Responses were generated using beam search to ensure stable and coherent outputs. **DLRG** (R and Rajalakshmi, 2026) employed a traditional machine learning pipeline based on TF-IDF vectorization and classical classifiers such as Linear Support Vector Classifier, Multinomial Naive Bayes, and Logistic Regression. These models were applied to represent textual features and generate predictions based on learned patterns in the training data.

Amritha team experimented with three different system configurations. The first run used a rule-based template matching approach combined with TF-IDF similarity to identify relevant training examples. The second run utilized the Llama-3.2-1B-Instruct model (Grattafiori et al., 2024) with chain-of-thought prompting and few-shot learning strategies (Brown et al., 2020). The third run used Google's Gemini model with multiple specialized prompt templates designed for different hate speech patterns and safety requirements. These approaches aimed to generate counter-narratives while prioritizing safety and contextual relevance.

[2] https://huggingface.co/Qwen/Qwen3.5-9B

Table 4: Scorecard of Task 1 - Span Detection. (Acc: Accuracy; mP: Macro Precision; mR: Macro Recall; mF1: Macro F1; wP: Weighted Precision; wR: Weighted Recall; wF1: Weighted F1)

Teams	Acc	mP	mR	mF1	wP	wR	wF1	Rank
English								
TeamV (Ulli and Kumari, 2026)	0.6354	0.5340	0.5396	**0.5338**	0.6674	0.6354	0.6493	1
DuoNova (S et al., 2026)	0.6494	0.5111	0.5110	**0.5111**	0.6487	0.6494	0.6490	2
Tamil								
TeamV	0.6624	0.5275	0.5270	**0.5272**	0.6591	0.6624	0.6607	1
DuoNova	0.7072	0.5247	0.5154	**0.5090**	0.6545	0.7072	0.6737	2
Hindi								
TeamV	0.5513	0.5486	0.5494	**0.5478**	0.5572	0.5513	0.5531	1
DuoNova	0.4648	0.4590	0.4585	**0.4585**	0.4684	0.4648	0.4663	2

SigJBS (Sinha et al., 2026) developed a two-stage pipeline combining progressive prompting and instruction fine-tuning. They used the Gemma 3 12B-IT model [3] optimized with Unsloth for efficient training using 4-bit quantization. The system initially explored multiple few-shot prompting configurations and later applied QLoRA-based supervised fine-tuning with LoRA adapters. Tamil data was oversampled to address language imbalance and improve generation quality.

JusticeBots (Pranesh et al., 2026) adopted a prompt-based approach using ChatGPT. Carefully designed prompts instructed the model to generate respectful and constructive responses while avoiding repetition of harmful language from the input comments. The system relied entirely on the instruction-following capabilities of the language model without additional training. **Team V** implemented a few-shot prompting strategy using the Qwen3-max model. A five-shot prompt containing examples from both English and Tamil guided the model to generate concise, empathetic counter-narratives in one to three sentences. The system ensured that responses were generated in the same language as the input. Post-processing steps were applied to remove model artifacts and extract the final response.

Overall, the participating teams explored a diverse range of approaches, including supervised transformer-based generation, prompt engineering with large language models, optimization-based prompting strategies, and classical machine learning methods. These methodologies highlight the evolving landscape of counter-speech generation techniques for addressing online hate speech in multilingual languages.

[3] https://huggingface.co/google/gemma-3-12b-it

6 Results and Discussion

For Task 1, the results summarized in Table 4 show that participating systems were able to achieve competitive performance across the three languages. Overall, the best-performing systems achieved macro-F1 scores of 0.5338 for English, 0.5272 for Tamil, and 0.5478 for Hindi. These results indicate that identifying precise hateful spans remains challenging due to linguistic variability, informal language usage, and the contextual nature of hate speech expressions. Across languages, models demonstrated relatively stable weighted F1 scores, suggesting that systems were generally effective at capturing the dominant classes despite dataset imbalance. However, the differences between macro and weighted metrics highlight the difficulty in consistently identifying all types of hateful expressions, particularly in low-resource scenarios such as Hindi, where the training data size is relatively small. Overall, the participating systems demonstrated strong capability in span-level detection using both supervised transformer-based models and prompt-based approaches.

For Task 2, Tables 5, and 6 present the performance of participating systems for English and Tamil. The results indicate that generating high-quality counter-narratives is a complex task that requires balancing semantic relevance, politeness, and contextual coherence. Systems achieved strong BERTScore values, typically above 85%, indicating that generated responses were semantically aligned with reference counter-narratives. However, greater variation was observed in rubric-based metrics such as Politeness and Respectfulness Score (PRS), Quality Score (QS), and Contextual Counter-Narrative Coherence (CCNC), reflecting differences in how well systems cap-

Table 5: Scorecard of Task 2 - Counter Narrative Generation for English (in %). (PRS: Politeness and Respectful Score; QS: Quality Score; CCNC: Contextual Counter-Narrative Coherence Score.)

Teams	Runs	Reference-Based Scores		Rubric-Based Scores			Overall Avg. (%)	Rank
		Distinct-2 (%)	BERTScore-F1 (%)	PRS (%)	QS (%)	CCNC (%)		
Team_V (Ulli and Kumari, 2026)	Run 1	73.56	88.78	90.91	90.15	93.94	87.47	1
SigJBS (Sinha et al., 2026)	Run 1	69.32	86.66	93.18	90.91	91.67	86.35	2
NEUNI (Gajawada et al., 2026)	Run 1	64.50	86.29	91.67	86.36	86.36	83.04	3
DLRG (R and Rajalakshmi, 2026)	Run 2	74.36	85.55	72.73	69.70	84.09	77.29	4
JusticeBots (Pranesh et al., 2026)	Run 1	79.11	87.63	76.52	52.27	57.58	70.62	5
RespectNLP (Priya and Bharathi, 2026)	Run 1	78.56	82.93	53.79	54.55	81.82	70.33	6
Amritha	Run 3	8.16	86.02	100.00	68.18	61.36	64.74	7
DuoNova (S et al., 2026)	Run 1	58.22	86.04	56.82	37.88	50.00	57.79	8

Table 6: Scorecard of Task 2 - Counter Narrative Generation for Tamil (in %). (PRS: Politeness and Respectful Score; QS: Quality Score; CCNC: Contextual Counter-Narrative Coherence Score.)

Teams	Runs	Reference-Based Scores		Rubric-Based Scores			Overall Avg. (%)	Rank
		Distinct-2 (%)	BERTScore-F1 (%)	PRS (%)	QS (%)	CCNC (%)		
DLRG (R and Rajalakshmi, 2026)	Run 3	27.30	85.73	100.00	97.71	91.28	80.40	1
Amritha	Run 3	20.89	85.27	100.00	100.00	89.45	79.12	2
NEUNI (Gajawada et al., 2026)	Run 2	19.16	85.09	95.41	86.24	92.66	75.71	3
JusticeBots (Pranesh et al., 2026)	Run 1	27.01	85.67	87.16	66.97	73.39	68.04	4
Team_V (Ulli and Kumari, 2026)	Run 1	25.61	86.25	87.61	55.50	66.51	64.30	5
SigJBS (Sinha et al., 2026)	Run 1	25.29	85.29	75.23	72.02	61.01	63.77	6
DuoNova (S et al., 2026)	Run 1	3.62	86.04	94.50	61.93	64.68	62.15	7
RespectNLP (Priya and Bharathi, 2026)	Run 1	17.43	80.23	50.92	11.47	7.80	33.57	8

tured the intended tone and contextual appropriateness of counter-speech. In English, the top systems achieved overall average scores above 85%, whereas in Tamil, the highest-performing systems reached approximately 80%. This difference highlights the additional challenges associated with generating high-quality responses in multilingual settings and for languages with comparatively fewer training resources.

Overall, the shared task demonstrates the complementary nature of detection and response generation approaches for addressing online hate speech. The span detection task highlights the importance of precise identification of harmful expressions, while the counter-narrative generation task emphasizes constructive mitigation strategies through respectful and contextually relevant responses. The results across both tasks indicate that transformer-based models and large language models are effective for these problems, although challenges remain in handling linguistic diversity, subtle forms of hate speech, and ensuring consistently high-quality counter-narratives. The shared task therefore provides a valuable benchmark for future research on multilingual hate speech detection and counter-speech generation.

7 Conclusion

This paper presented the findings of the shared task on generating counter-narratives for homophobic and transphobic comments. The task consisted of two subtasks: span detection and counter-narrative generation. The results demonstrate that transformer-based models and large language models can effectively detect hateful spans and generate constructive responses to harmful content. In the span detection task, participating systems achieved competitive performance across English, Tamil, and Hindi using a combination of fine-tuned transformer models and prompt-based approaches. For the counter-narrative generation task, the models produced responses with strong semantic alignment to reference counter-narratives, as reflected in high BERTScore values. However, rubric-based evaluation assessing politeness, quality, and contextual coherence revealed variations in system performance. The results highlight the challenges associated with multilingual and low-resource language settings, particularly in generating contextually appropriate counter-speech. Overall, this shared task provides a useful benchmark for studying both the detection of harmful expressions and the generation of constructive counter-narratives. Future work can focus on expanding multilingual datasets, improving evaluation frameworks, and exploring human-in-the-loop approaches to ensure more reliable and socially responsible counter-narrative generation systems.

8 Ethical Considerations

The shared task deals with the sensitive topic of homophobia and transphobia, which may involve

the use of derogatory language. While this is required for the task of hate speech detection and counter-speech generation, it may be distressing for developers and for annotators. The focus here is on the dataset's use for scientific research and we advise against using the dataset for any other purpose. The aim of the systems developed in this task is to support respectful counter-narrative generation by providing constructive alternatives to avoid the spread of hate speech, while refraining from spreading further the original, potentially offensive language. However, even with the best efforts, there is always a risk that the outcome of such systems can be unpredictable and, hence, potentially harmful, especially if the training data is not sufficient or if there are biases in the systems. Therefore, outputs generated by such systems should be carefully reviewed before being released into the wild. Models developed in this shared task are intended to be used to foster safe and inclusive online environments, in alignment with the principles of safe AI, diversity, and languages.

Acknowledgement

This work was conducted with the financial support from Research Ireland under Grant Number SFI/12/RC/2289_P2(Insight_2), Research Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223; and a grant from the College of Science and Engineering, University of Galway, Ireland.

References

Susan Benesch. 2014. Countering dangerous speech: New ideas for genocide prevention. *Available at SSRN 3686876*.

Matthew N Berger, Melody Taba, Jennifer L Marino, Megan SC Lim, and S Rachel Skinner. 2022. Social Media use and Health and Well-being of Lesbian, Gay, Bisexual, Transgender, and Queer Youth: Systematic Review. *Journal of medical Internet research*, 24(9):e38449.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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, Prasanna Kumar Kumaresan, Ruba Priyadharshini, Paul Buitelaar, Asha Hegde, Hosahalli Lakshmaiah Shashirekha, Saranya Rajiakodi, Miguel Ángel García-Cumbreras, Salud María Jiménez-Zafra, José Antonio García-Díaz, Rafael Valencia-García, Kishore Kumar 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, Rahul Ponnusamy, Malliga Subramanian, Paul Buitelaar, Miguel Ángel García-Cumbreras, Salud María Jiménez-Zafra, José Antonio García-Díaz, Rafael Valencia-García, and Nitesh Jindal. 2023. "overview of second shared task on homophobia and transphobia detection in social media comments". In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 38–46, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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

Cheng-Han Chiang and Hung-yi Lee. 2023. "can large language models be an alternative to human evaluations?". In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tai, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2024. Scaling Instruction-Finetuned Language Models. *J. Mach. Learn. Res.*, 25(1).

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. "towards knowledge-grounded counter narrative generation for hate speech". In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. "unsupervised cross-lingual representation learning at scale". In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Sander De Ridder and Sofie Van Bauwel. 2015. The Discursive Construction of Gay Teenagers in Times of Mediatization: Youth's Reflections on Intimate Storytelling, Queer Shame and Realness in Popular Social Media Places. *Journal of Youth Studies*, 18(6):777–793.

María José Díaz-Torres, Paulina Alejandra Morán-Méndez, Luis Villaseñor-Pineda, Manuel Montes, Juan Aguilera, and Luis Meneses-Lerín. 2020. Automatic detection of offensive language in social media: Defining linguistic criteria to build a Mexican Spanish dataset. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 132–136.

Mekselina Doğanç and Ilia Markov. 2023. From Generic to Personalized: Investigating Strategies for Generating Targeted Counter Narratives against Hate Speech. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 1–12, Prague, Czechia. Association for Computational Linguistics.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3226–3240, Online. Association for Computational Linguistics.

Preethi Gajawada, Bhanu Harsha Yanamadala, Akankshya Kar, Sahil Wadhwa, and Divya Chaudhary. 2026. Neuni@lt-edi 2026: Counter narrative generation on homophobic and transphobic comments. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Zhiwei Gao, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2020. Offensive language detection on video live streaming chat. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1936–1940.

Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2020. Countering hate on social media: Large scale classification of hate and counter speech. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 102–112, Online. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Darryl B Hill. 2002. Genderism, Transphobia, and Gender Bashing: A Framework for Interpreting Anti-Transgender Violence. *Understanding and dealing with violence: A multicultural approach*, 4:113–137.

Rachel Keighley. 2022. Hate hurts: Exploring the Impact of online hate on LGBTQ+ Young People. *Women & Criminal Justice*, 32(1-2):29–48.

Prasanna Kumar Kumaresan, Bharathi Raja Chakravarthi, Ruba Priyadharshini, Paul Buitelaar, Malliga Subramanian, and Kishore Kumar Ponnusamy. 2025a. "overview of homophobia and transphobia span detection in social media comments". In *Proceedings of the 5th Conference on Language, Data and Knowledge: Fifth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 229–234, Naples, Italy. Unior Press.

Prasanna Kumar Kumaresan, Devendra Deepak Kayande, Ruba Priyadharshini, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2025b. Homophobia and Transphobia Span Identification in Low-resource Languages. *Natural Language Processing Journal*, page 100169.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. "a diversity-promoting objective function for neural conversation models". In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

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.

Jody Norton. 1997. "Brain says you're a girl, but I think you're a sissy boy": Cultural origins of Transphobia. *International Journal of Sexuality and Gender Studies*, 2:139–164.

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

TT Pranesh, KK Thamizhmathi, S Vigneshwaran, and B Bharathi. 2026. Justicebots@lt-edi 2026: Prompt-based counter-narrative generation for homophobia and transphobia comments. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Praveen Prasannan, Prasanna Kumar Kumaresan, Saranya Rajiakodi, CN Subalalitha, and Bharathi Raja Chakravarthi. 2025. Counter-Speech Generation for Homophobic and Transphobic Social Media Content in Malayalam. *Social Network Analysis and Mining*, 15(1):87.

S.B. Priya and B Bharathi. 2026. Rspectnlp@lt-edi 2026:rubric-driven prompting for safe multilingual counter narrative generation. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ramesh Kannan R and Ratnavel Rajalakshmi. 2026. Dlrg@lt-edi 2026: Automating counter-narratives for homophobic and transphobic comments. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Tharindu Ranasinghe and Marcos Zampieri. 2020. "multilingual offensive language identification with cross-lingual embeddings". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online. Association for Computational Linguistics.

Manasa S, Arohi Rawat, and Anbukkarasi S. 2026. Duonova@ltedi 2026: Multilingual span detection and counter-narrative generation on homophobic and transphobic comments. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Siva Sai and Yashvardhan Sharma. 2021. Towards Offensive Language Identification for Dravidian Languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 18–27.

Robert D Schope and Michele J Eliason. 2004. Sissies and Tomboys: Gender Role Behaviors and Homophobia. *Journal of Gay & Lesbian Social Services*, 16(2):73–97.

Gaurangi Sinha, Rajarajeswari Palacharla, and Manoj Balaji Jagadeeshan. 2026. Sigjbs@lt-edi 2026: Qlora-tuned homophobic and transphobic counter narrative generation. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. "generating counter narratives against online hate speech: Data and strategies". In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.

Vinay Babu Ulli and Jyoti Kumari. 2026. Teamv at lt-edi 2026: Multilingual hate speech span detection and counter-narrative generation via few-shot in-context learning. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Fikri Aji. 2023. "multilingual large language models are not (yet) code-switchers". In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12567–12582, Singapore. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. *arXiv preprint arXiv:1904.09675*.