DuoNova@LTEDI 2026: Multilingual Span Detection and Counter-Narrative Generation on Homophobic and Transphobic Comments

Manasa S and Arohi Rawat and Anbukkarasi S
Manipal Institute of Technology Bengaluru,
Manipal Academy of Higher Education, Manipal,India
manasasudhar@gmail.com
arohi.rawat454@gmail.com
anbukkarasi.s@manipal.edu

Abstract

The detection and response to homophobic and transphobic comments are important challenges in Natural Language Processing. In this paper, we focus on the detection of span for homophobic and transphobic comments (Task 1) and generation of counter narratives for abusive comments (Task 2) for the LT-EDI @ ACL 2026 shared task. Harmful comments made online against the LGBTQ+ community have created a hostile environment for users. In this paper, we have used the transformer model for the detection of span for homophobic and transphobic comments and generation of counter narratives. In this task, the detection of the span of comments containing homophobic and transphobic words and the generation of counter narratives for abusive comments have been done using the transformer model. The results show the efficiency of the transformer model in the detection of the span of comments and generation of counter narratives. This paper emphasizes the efficiency of the transformer model in creating a safe environment for users.

1 Introduction

With the rapid increase in the number of social media platforms, the way individuals communicate and express their opinions online has also changed. However, the rapid growth of social media has also increased the amount of harmful content and discriminatory behavior, especially with reference to the LGBTQ+ community. Such homophobic and transphobic content creates a hostile online environment. Thus, handling such content is a significant challenge for Natural Language Processing (NLP) systems.

Most NLP systems currently only deal with the removal of harmful content. However, the detection is not sufficient; it does not encourage constructive engagement. It is also significant to identify the exact span of the text where the harmful content is present. It will be more interpretable. However, the generation of counter-narratives is another way of handling harmful content. It is a way of responding to harmful content with respectful and empathetic content. It does not remove the engagement; rather, it decreases the hostility and encourages constructive engagement.

The development of effective span detection and counter-narrative generation is a complex task. The model should understand the context of the hate comment, the specific hate content, and generate appropriate responses without being offensive. This is a more complex problem in a multilingual setting, where linguistic and cultural variations are more common between languages.

In this paper, we are participating in the shared task of LT-EDI @ ACL 2026. This shared task consists of two tasks: Task 1 is homophobia and transphobia span detection, and Task 2 is counter-narrative generation. For Task 1, we are proposing a token classification model using a transformer-based model for the detection of homophobia and transphobia. For Task 2, we are proposing a sequence-to-sequence model using a transformer-based model. The rest of the paper is as follows: In Section 2, we discuss the related work. In Section 3, we discuss the dataset. In Section 4, we discuss the methodology. In Section 5, we discuss the error analysis. In Section 6, we showcase our experimental result.In section 7 we conclude the paper.

2 Related Work

The detection of hate speech and homophobia has been studied thoroughly in recent years (Chakravarthi, 2024). Recent research has also been carried out on the identification of homophobic and transphobic content at the span level, especially in low-resource languages (Kumaresan et al., 2025). Another research area that has been explored recently is the generation of counter-speech,

Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 167–171

which is seen as an alternative solution to the removal of hateful content from online platforms (Prasannan et al., 2025). Some other studies have explored hate speech detection and the spread of toxic content on social media, as well as the development of counter-narrative datasets such as CONAN (Mozafari et al., 2019; Mathew et al., 2019; Chung et al., 2019).

Hate speech detection has been researched in various languages, including Dravidian languages. Various shared tasks and datasets have been proposed for the research of hate speech in Dravidian languages (Chakravarthi et al., 2021; Priyadharshini and Chakravarthi, 2021). Past research has also been carried out on the identification of offensive content in Dravidian languages using a multilingual approach (Chakravarthi et al., 2020).

Significant progress has been achieved in the development of transformer-based models, which has led to improved performances in various natural language processing tasks, such as hate speech detection and text generation. BERT and T5 models have shown remarkable capabilities in learning contextual representations and generating coherent text (Devlin et al., 2019; Raffel et al., 2020).

This study extends the recent developments in transformer-based models and proposes a framework that incorporates the transformer model to solve the tasks of span detection and counter narrative generation in English and Tamil languages. The study focuses on the identification of homophobic and transphobic spans in online comments and the generation of appropriate counter narratives, which are part of the LT-EDI shared task.

3 Dataset Description

The dataset was provided as part of the LT-EDI @ ACL 2026 shared task (Kumaresan et al., 2026)., covering span detection (Task 1) and counter-narrative generation (Task 2). The training data comprises 1,800 instances, and the test data comprise 66 instances. For each training instance, the data include an id, the original user comment (text), the annotated harmful span (span), the corresponding human-written counter narrative (counter_narrative), and a label for abusive content.

For the generation task, the entire comment text is employed as the input, and the corresponding counter narrative is employed as the target output. Though the dataset offers span-level annotations that focus on harmful content, the entire comment is employed for the task. The dataset exhibits class imbalance, with non-harmful tokens significantly outnumbering harmful tokens, which may affect model performance. The test data only includes the comment text, and the task for the system is to generate counter narratives for this unseen input.

To evaluate the performance of the model during training, a validation split was created from the training data.

Before fine-tuning, the text data was cleaned and tokenized using the tokenizer of the chosen pre-trained transformer model.

4 Proposed Methodology

This section presents a multilingual transformer-based system for the detection of homophobia and transphobia span and generation of counter-narrative. We formulate span detection as a token classification task using the BIO tagging scheme, where tokens are labeled as B (beginning), I (inside), or O (outside) of a harmful span. The model is trained using cross-entropy loss over token-level predictions. Class imbalance is not explicitly handled through class weighting.

Predicted token labels are converted into spans by grouping consecutive tokens labeled as B and I. Post-processing is applied to merge subword tokens into complete words.

The input to the model follows a prompt-based format: "Generate a polite and respectful counter-narrative for the following comment: <input>". Decoding is performed using greedy decoding with a maximum sequence length of 128. No additional constraints were explicitly applied to filter harmful or off-topic outputs, as the model relies on its pretrained instruction-following capabilities.

The proposed system uses a pre-trained transformer-based model, FLAN-T5 (Raffel et al., 2020) for English and mT5 (Xue et al., 2021) for Tamil and Hindi.For span detection, a token classification head is used on top of the encoder outputs, while FLAN-T5 is used in a generative manner for counter-narrative generation. While the approach follows a standard transformer-based pipeline, this work demonstrates the effectiveness of a unified multilingual framework for both span detection and counter-narrative generation across multiple languages.

4.1 Pre-Processing

The input dataset for this task comprises comments obtained from social media, along with corresponding span indices. The text is tokenized using the pre-trained tokenizer of the selected transformer model. Offset information is used to map each token to its original character position, enabling token-level labeling for span detection. Padding and truncation are applied to ensure a maximum sequence length of 128. The complete system pipeline involves:

Figure 1: pipeline

This method utilizes the pretrained transformer model for effective multilingual hate speech span detection and counter-narrative generation.

4.2 English Span Detection and Counter-Narrative Generation using FLAN-T5

For English, the google/flan-t5-base model is used for Task 1 and Task 2. FLAN-T5 is an instruction-tuned transformer model that can comprehend and create context-aware text. For span detection, the model learns the contextual meaning of the input text and detects the harmful parts. For counter-narrative generation, the model uses prompt-based input to generate empathetic and respectful content. The parameters used for the model are shown below:

Parameter	Value
Model	google/flan-t5-base
Maximum sequence length	128
Batch size	8
Epochs	3
Learning rate	5e-5
Framework	HuggingFace Transformers

Table 1: Parameters used in FLAN-T5 Model

The model makes predictions for English span detection and counter-narratives.

4.3 Tamil and Hindi Span Detection and Counter-Narrative Generation using mT5

For Tamil Task 1, Tamil Task 2, and Hindi Task 1, the pre-trained google/mt5-small model is used. mT5 is a multilingual transformer model that can handle multiple languages. Tokenization and embedding generation are performed using the mT5 tokenizer. For span detection, a token classification head is used. For counter-narrative generation, sequence generation is applied. The parameters used for the model are shown in Table 2.

Parameter	Value
Model	google/mt5-small
Maximum sequence length	128
Batch size	8
Epochs	3
Learning rate	5e-5
Framework	HuggingFace Transformers

Table 2: Parameters used in mT5 Model

The model predicts hateful spans and generates counter-narratives for Tamil and Hindi datasets.

5 Experimental Results

The proposed system was assessed in the LT-EDI @ ACL 2026 shared task for span detection (Task 1) and counter narrative generation (Task 2). For reporting the experimental results, the official rank lists and evaluation metrics provided by the organizers are used.

5.1 Task 1: Span Detection

For the span detection task, the system was assessed in terms of Accuracy, macro Precision (mP), macro Recall (mR), macro F1 (mF1), weighted Precision (wP), weighted Recall (wR), and weighted F1 (wF1) metrics. For the rank list, the primary metric was macro F1. The proposed system achieved second rank for English and Tamil corpora. For English and Tamil corpora, the weighted F1 metrics were 0.6490 and 0.6737. In the case of the Hindi language, the performance was relatively lower, and this might be due to the limited availability of data and linguistic variations. Nevertheless, the system was ranked second for all three languages.

Language	Macro F1 Score
English	0.5111
Tamil	0.5090
Hindi	0.4585

Table 3: Macro F1 scores for Task 1: Span Detection across languages.

The relatively lower macro F1 scores may be influenced by class imbalance in the dataset, which makes accurate span identification more challenging.

5.2 Task 2: Counter-Narrative Generation

In the case of the counter-narrative generation, the evaluation was carried out using a combination of reference-based metrics, such as Distinct-2 and BERTScore-F1, and rubric-based metrics, such as Politeness, Respectful Score (PRS), Quality Score(QS), and Contextual Counter-Narrative Coherence(CCNC). The ranking was carried out based on the overall average score. Task 2 was defined only for English and Tamil in the shared task; therefore, results for Hindi are not reported.

5.2.1 English

In English, the system received competitive reference-based scores and moderate rubric-based scores. It scored a BERTScore-F1 of 86.04% and a general average score of 57.79%.

5.2.2 Tamil

In Tamil, the system achieved strong politeness and respectful behavior with a PRS score of 94.50%. The overall average score was 62.15%. The high PRS score indicates that the generated responses were consistently polite and respectful.

The performance difference between English and Tamil suggests that the model generates more contextually appropriate and polite responses in Tamil, while English responses tend to be more generic.

Metric	English (%)	Tamil (%)
Distinct-2	58.22	3.62
BERTScore-F1	86.04	86.04
PRS	56.82	94.50
QS	37.88	61.93
CCNC	50.00	64.68
Overall Average	57.79	62.15

Table 4: Task 2: Counter-Narrative Generation Results for English and Tamil.

The lower rubric-based scores in English may be due to the model generating more generic responses, whereas Tamil outputs tend to be more consistent due to simpler sentence structures and lower linguistic variability. Due to the constraints of the shared task setting, we did not evaluate additional baseline models or perform ablation studies. The observed performance can be primarily attributed to the strong pretrained capabilities of FLAN-T5 and mT5 models, which are effective in both understanding contextual information and generating coherent responses. Future work will include comparisons with simpler baselines and ablation studies to better understand the contribution of different components.

6 Error Analysis

The proposed system was effective in producing meaningful counter-narratives in most cases. Nevertheless, a number of errors were identified during evaluation.

One of the major issues observed was grammatical inconsistency, particularly in the Tamil outputs generated using the mT5 model. In a few instances, there were minor grammatical errors or improper sentence formation, which can be attributed to the limitations of the pretrained multilingual model.

Another issue was the generation of incomplete or generic responses. In some cases, the model produced very short responses that were not effective in addressing the input comment. This was more prevalent when the comment was complex or ambiguous. Additionally, certain responses were not context-specific and instead consisted of general statements about respect and inclusion. This behavior may be due to the model relying on general patterns learned during training.

In the case of span detection, errors were observed when the harmful content was implicit or context-dependent. For example, the model often failed to identify spans in cases where the harmful intent was implied rather than explicitly stated.

Furthermore, the relatively small size of the dataset may have limited the model's ability to generalize across diverse linguistic patterns, impacting both span detection accuracy and the quality of generated responses. It also introduces a risk of overfitting, where the model may learn dataset-specific patterns rather than generalizable features. Additionally, we did not perform multiple training runs to evaluate variance, and therefore the stability

of the results across different random initializations is not explicitly measured. This remains an area for future investigation.

Code Availability

The source code is publicly available at:https://github.com/Manasa-S-02/DuoNova.git.

7 Conclusion

This paper presented a transformer-based approach for homophobia and transphobia span detection and counter-narrative generation as part of the LT-EDI @ ACL 2026 shared task. For span detection (Task 1), the FLAN-T5 model was utilized for English, and the mT5 model was utilized for Tamil and Hindi. For counter narrative generation (Task 2), the FLAN-T5 model was utilized for English and the mT5 model was utilized for Tamil. Experimental results showed the effectiveness of the proposed method in detecting homophobia and transphobia and generating polite and contextually appropriate counter narratives.

References

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, Ruba Priyadharshini, Vigneshwaran Muralidaran, Elizabeth Sherly, and John McCrae. 2021. Overview of the dravidian-codemix 2021 shared task on offensive language identification. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, and Elizabeth Sherly. 2020. Multilingual offensive language identification in dravidian languages. *Information Processing and Management*.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroğlu, and Marco Guerini. 2019. Conan – counter narratives through nichesourcing. In *Proceedings of ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

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*, page 100169.

Prasanna Kumar Kumaresan, Praveen Prasannan, Tanay Singh, Ruba Priyadharshini, Subalalitha Chinnaudayar Navaneethakrishnan, Saranya Rajiakodi, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2026. Findings of the shared task on 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.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the Web Conference*.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A bert-based transfer learning approach for hate speech detection. In *International Conference on Complex Networks*.

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.

Ruba Priyadharshini and Bharathi Raja Chakravarthi. 2021. Dravidian offensive language dataset in tamil, malayalam, and kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of NAACL*.