SKV trio@LT-EDI-2025: Hybrid TF-IDF and BERT Embeddings for Multilingual Homophobia and Transphobia Detection in Social Media Comments

Konkimalla Laxmi Vignesh [1], Mahankali Sri Ram Krishna[1], Dondluru Keerthana[1], Premjith B[1],
[1]Amrita School of Artificial Intelligence, Coimbatore,
Amrita Vishwa Vidyapeetham, India
Correspondence: b_premjith@cb.amrita.edu

Abstract

This paper presents a description of the paper submitted to the Shared Task on Homophobia and Transphobia Detection in Social Media Comments, LT-EDI at LDK 2025. We propose a hybrid approach to detect homophobic and transphobic content in low-resource languages using Term Frequency-Inverse Document Frequency (TF-IDF) and Bidirectional Encoder Representations from Transformers (BERT) for contextual embeddings. The TF-IDF helps capture the token's importance, whereas BERT generates contextualized embeddings. This hybridization subsequently generates an embedding that contains statistical surface-level patterns and deep semantic understanding. The system uses principal component analysis (PCA) and a random forest classifier. The application of PCA converts a sparse, very high-dimensional embedding into a dense representation by keeping only the most relevant features. The model achieved robust performance across eight Indian languages, with the highest accuracy in Hindi. However, lower performance in Marathi highlights challenges in low-resource settings. Combining TF-IDF and BERT embeddings leads to better classification results, showing the benefits of integrating simple and complex language models. Limitations include potential feature redundancy and poor performance in languages with complex word forms, indicating a need for future adjustments to support multiple languages and address imbalances.

1 Introduction

Homophobia and transphobia are two concepts that foster negative opinions about homosexual and transgender individuals (Chakravarthi et al., 2022). These concepts are endemic in online communities and are most commonly expressed through context-dependent, subtle language that excludes LGBTQ+ populations. As digital communication grows, automated systems to detect and mitigate negative content using discriminatory language become increasingly important. However, detecting such content in low-resource languages is a significant challenge posed by linguistic subtleties and the lack of available annotated data.

This paper addresses the submission of the team SKV trio to the shared task on Homophobia and Transphobia Detection in Social Media Comments LT-EDI at LDK 2025 (Kumaresan et al., 2025). The proposed system was designed as a classification model with hybrid features. The proposed model combines features obtained from Term Frequency-Inverse Document Frequency (TF-IDF) and a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model. Integrating these categories of features, the system achieves both surface patterns and deeper contextual meaning required for efficient classification. The classifier was trained using a random forest classifier with the combined feature representations. We employed Principal Component Analysis (PCA) to facilitate the reduction of dimensionality within the TF-IDF embedding to match the TF-IDF and BERT embedding dimensionality. Our experiments demonstrate that the fusion of TF-IDF and BERT embeddings significantly improves classification accuracy compared to using either feature extraction technique alone. The implementation is available at https://github.com/K-LAXMIVIGNESH/SKVtrio-/tree/main

2 Related Work

The study (Chakravarthi et al., 2022) to detect homophobia and transphobia makes the hypothesis that multilingual language models can detect hate speech on social media more effectively if data augmentation through pseudolabeling is used. To test the hypothesis, the study looked at a number of multilingual language models. With 5,193 Malayalam and 3,203 Hindi comments, Kumare-

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 26–30

san P et al. (Kumaresan et al., 2023) offer a new, high-quality dataset for identifying homophobic and transphobic content in both languages. The dataset also contains tests on the Malayalam, Hindi, English, Tamil, and Tamil-English datasets using transformer-based deep learning models and conventional machine learning. To automatically identify homophobic and transphobic content, (Chakravarthi, 2024) proposed a new dataset that was annotated by experts and trained machine learning models using it. 15,141 annotated comments in Tamil, English, and both languages are included in the dataset. The average macro F1 score for the top Tamil, English, and Tamil-English systems was 0.570, 0.870, and 0.610, respectively. These scores were obtained by training a random forest with fastText for Tamil data and logistic regression with BERT features for English and Tamil-English data. (Kumaresan et al., 2024) proposed a new expert-labeled dataset for Telugu, Kannada, and Gujarati, along with an expert-labeled dataset. The dataset includes extensive annotation rules, gathering about 10,000 comments in all three languages. A baseline model with pre-trained transformers was trained using the proposed dataset. The findings of a shared task to detect homophobic and transphobic language in social media posts are presented in (Chakravarthi et al., 2023) and (Chakravarthi et al., 2024). Task B required a fine-grained seven-class classification, while Task A classified comments into homophobic, transphobic, or neither categories. English, Spanish, Tamil, Hindi, and Malayalam were the five languages in which the task contained data. With the highest F1-scores of 0.997 for Spanish in Task A and 0.884 for Tamil in Task B, the top systems showed excellent performance. All the models reported in the literature are using either conventional techniques or BERT-based models for generating features. BERT-based models may not generate relevant features for rare words, which are important. These words can be efficiently captured by the TF-IDF algorithm.

3 Dataset description

The dataset is divided into training, development, and testing sets for each language. Table 1 shows the number of samples for each split.

There is noticeable variation in dataset size across languages. Kannada and Telugu, for instance, have significantly more samples than Hindi or Tamil, introducing language imbalance that can impact model generalization.

Table 1: Data distribution across languages

Language	Train	Dev	Test
English	3164	792	990
Gujarati	8119	1740	1740
Hindi	2560	320	321
Kannada	10063	2157	2156
Malayalam	3114	1213	866
Marathi	3500	750	750
Tamil	2662	666	833
Telugu	9050	1940	1939

4 Methodology

This section outlines the complete methodology of the system designed for this task. The methodology is divided into the following stages: data preprocessing, feature extraction, embedding fusion, model training, and prediction. Figure 1 illustrates the machine learning pipeline we used for this shared task. Three different data sources support the model: training, development, and test data.

In the preprocessing step, we check for blank rows and columns and remove such rows and columns from the dataset. We also searched for and removed any rows and columns that lacked text or labels. Finally, we converted all the labels into integers for processing.

We passed the preprocessed data through two different paths to extract two different types of features. We used the TF-IDF and BERT algorithms for feature extraction. TF-IDF captures the importance of words and tokens based on their frequency in the corpus. The TF-IDF feature extraction module was followed by an PCA module for dimensionality reduction. This algorithm uses 42 principal components for sampling features from a relatively lower dimensional space. BERT extracts contextual information from the input sequence and transforms the input text into a vector representation. Here, we used MuRIL (Sreelakshmi et al., 2024) for generating the embeddings. MuRIL is trained over translated and transliterated data in addition to the text in the original script. This motivated us to use MuRIL embeddings for this task.

We combine these two dense and lower-dimensional representations in a feature fusion step. We concatenated the contextual embeddings and dimensionality-reduced TF-IDF embeddings in this

step. We performed a 5-fold cross-validation to ensure the model's robustness. Later, we trained the model using the training data and evaluated its performance using the development data. Finally, we used the model to predict the labels of the test data.

5 Experimental Results

We evaluated our hybrid model on eight Indian languages using average F1-score and accuracy as the primary metrics. The results indicate that the model performs consistently well across most languages, especially in high-resource conditions.

The best performance was observed for Hindi, followed by English and Malayalam. These results suggest that the model effectively captures linguistic features and contextual nuances in these languages.

Table 2: Average accuracy and F1-score across languages computed for validation data

Language	Accuracy	F1-score
Hindi	0.9465	0.9205
English	0.9425	0.9157
Malayalam	0.9249	0.9187
Gujarati	0.9018	0.9022
Telugu	0.8977	0.8981
Tamil	0.8674	0.8442
Kannada	0.8587	0.8593
Marathi	0.7394	0.6340

The performance of the proposed system on the test data is shown in Figure 3.

Table 3: Average accuracy and F1-score across languages computed for test data

Language	F1-score
Hindi	0.3260
English	0.3350
Malayalam	0.3960
Gujarati	0.8570
Telugu	0.8660
Tamil	0.3710
Kannada	0.8120
Marathi	0.2940

From the results, it is evident that the model overfits on all languages except Gujarati, Kannada, and Telugu, showing strong generalization capabilities, even in moderately resourced or code-mixed environments. Marathi had the lowest F1-score among all languages, indicating possible challenges such as limited data, quality issues, or class imbalance. Kannada showed slightly lower results compared to others, but still maintained a reasonable performance.

Overall, the combination of TF-IDF and BERT embeddings has proven effective for multilingual hate speech classification. These results highlight the need for further work on handling dataset imbalance and enhancing model performance for low-resource languages.

The hyperparameters used for building the model are listed in Table 4.

Hyperparameters	Value
TF-IDF max_features	5000
dim(reduced TF-IDF)	512
RBF Sampler gamma	1.0
BERT truncation	True
BERT padding	True
BERT max_length	512
Random forest n_estimators	100
Random forest criterion	Gini
Random forest max_features	sqrt

Table 4: Hyperparameters used for building the model

6 Conclusion

In this work, we present a hybrid method that combines TF-IDF feature extraction along with BERT-based contextual embeddings for multilingual transphobia and homophobia detection in Indian languages. Our work demonstrates the effectiveness of combining both shallow lexical and deep contextual representations to achieve efficient classification. The experimental outcomes indicate that the hybrid model is both strong at capturing surface patterns and deeper semantic meanings required to detect subtle hate speech.

7 Limitations

The fusion of TF-IDF and BERT features can lead to feature redundancy or overfitting, as TF-IDF captures frequency-based semantics while BERT captures context, potentially introducing conflicting signals. We used the general BERT model for feature extraction, which may cause underperformance on low-resource or morphologically rich languages, especially if no multilingual fine-tuning is done.

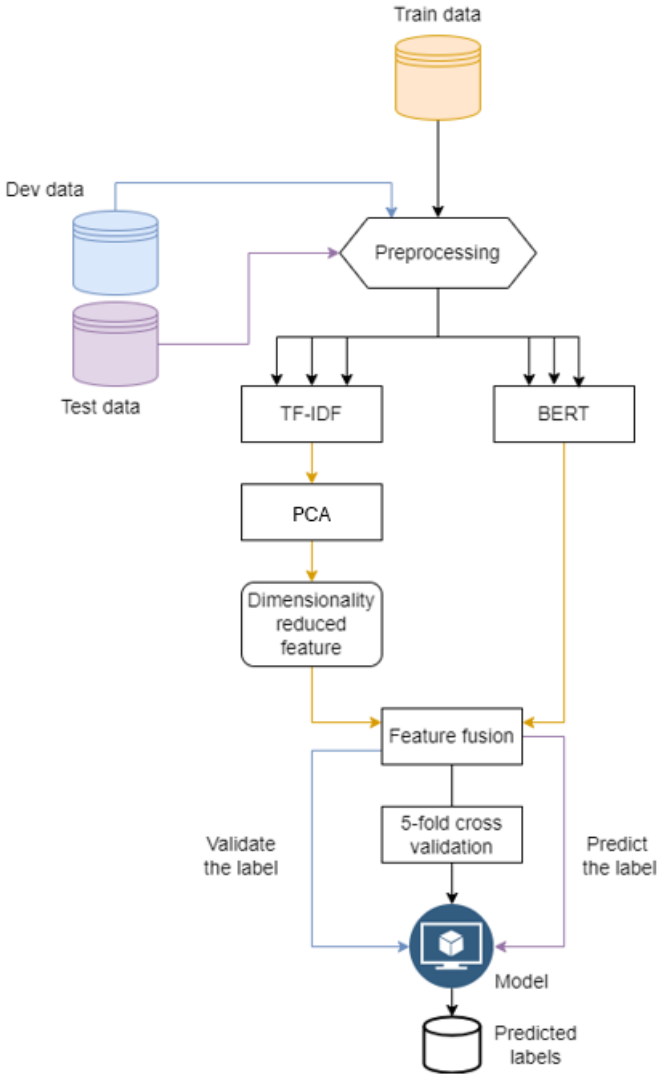

Figure 1: The block diagram explaining the workflow

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, Adeep Hande, Rahul Ponnusamy, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022. 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, Prasanna Kumaresan, Ruba Priyadharshini, Paul Buitelaar, Asha Hegde, Hosahalli Shashirekha, Saranya Rajiakodi, Miguel Ángel García, Salud María Jiménez-Zafra, José García-Díaz, and 1 others. 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.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Paul Buitelaar, Miguel Ángel García-Cumbreras, Salud María Jiménez-Zafra, José Antonio García-Díaz, Rafael Valencia-García, Nitesh Jindal, and 1 others. 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.

Prasanna Kumar Kumaresan, Bharathi Raja Chakravarthi, Ruba Priyadharshini, Paul Buitelaar, Malliga Subramanian, and Kishore Kumar Ponnusamy. 2025. Overview of fourth shared task on homophobia and transphobia span 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).

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.

K Sreelakshmi, B Premjith, Bharathi Raja Chakravarthi, and KP Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*.