SCaLAR@LT-EDI-2024: Homophobia/Transphobia Detection for Equality, Diversity, and Inclusion

Anonymous ACL submission

Abstract

Online consumers are badly impacted by the spread of harmful content on social media platforms. Content marked as homophobic or transphobic denotes hateful remarks to lesbian, gay, bisexual, or transgender individuals. It results in disrespectful statements and serious societal issues that can poison online spaces for LGBT+ people and work to eradicate equality, diversity, and inclusion. Our classification system in this research predicts, given remarks, whether or not they contain any homophobia or transphobia. The suggested models for the English language placed fourth with an outstanding F1-score of 0.438, other suggested model for Telugu received a weighted F1-score of 0.911 and was ranked fifth, for Hindi received a weighted F1-score of 0.410 and was ranked second, for Kannada, it received a weighted F1-score of 0.903 and was ranked seventh.

1 Introduction and Related Work

Homophobic/Transphobic [Diefendorf and Bridges, 2020] content on social media seeks to harm individuals who identify as lesbian, gay, or bisexual (LGB) through derogatory labels and denigrating phrases [Szymanski et al., 2008].

This abuse can manifest in physical violence, such as assault or sexual violence, and invasion of privacy by exposing personal information.

The rise of social media platforms has allowed individuals to express their thoughts and opinions [Gkotsis et al., 2016]. Still, it has also enabled the spread of unpleasant and offensive content [Zampieri et al., 2019].

The proliferation of offensive language online is a global phenomenon observed across various social media platforms like Facebook, YouTube, and Twitter [Gao et al., 2020].

This trend is particularly distressing for vulnerable groups such as lesbian, gay, bisexual, transgender, and other (LGBT+) individuals [Díaz-Torres et al., 2020]. Discrimination and violence against LGBT+ individuals are prevalent worldwide, violating their rights and subjecting them to torture and even execution [Barrientos et al., 2010]. It is important to recognize that sexual orientation and gender identity are integral parts of one's identity and should never be a basis for discrimination or abuse [Thurlow, 2001]. Many LGBT+ individuals turn to social media to seek support and connect with like-minded individuals as they face significant risks in their offline lives [Adkins et al., 2018]. Identifying and addressing homophobic/transphobic content on social media platforms is crucial to combat this societal problem, promoting equality, diversity, and inclusion and creating safer online spaces for LGBT+ individuals.

However, existing regulations often fail to provide adequate protection based on gender identity or expression, particularly for transgender adolescents [McGuire et al., 2010].

	Train	Dev	Test
English	3164	792	990
Hindi	2560	320	425
Telugu	9050	1940	1939
Kannada	10063	2157	2156

Table 1: Dataset Statics for training, development and test sets for English, Hindi, Telugu and Kannada

2 Shared Task Description

Participants are given remarks pulled from social media for the collaborative assignment. Predicting whether or not it contains any homophobia or transphobia detection was a challenge. Seed data [Chakravarthi et al., 2021], sampled as shown in Table 1, is given to the participants. The content is manually marked in the comments to indicate whether it contains homophobic or transphobic language. For all the languages in Table 2, we also conducted reports on data distribution among non-anti-LGBT+, homophobic, and transphobic mate-

	Language-wise distribution (Train + Dev)			
Label	**English**	**Hindi**	**Telugu**	**Kannada**
Non-anti-LGBT+ content	3726	2728	4243	5418
Homophobic	9	47	3235	3350
Transphobic	221	105	3512	3452

Table 2: Data distribution for the Homophobia/Transphobia Detection in social media comments database.

Comment	Label
I support her, very smart ponnu	Non-anti-LGBT+ content
Stupid film there is no gays in the world these are all their imagine only	Homophobic
Hey seriously I thought She was a Transgender	Transphobic

Table 3: Examples for Non-anti-LGBT+ content, Homophobic, Transphobic in the Homophobia/Transphobia Detection in social media comments dataset.

rials. Table 3 lists several instances of comments that are homophobic, transphobic, and non-anti-LGBT+.

3 Proposed Methodology

3.1 Overview

The study aims to develop robust systems for detecting Homophobia and Transphobia in social media comments across English, Hindi, Kannada, and Telugu languages. Three distinct systems have been designed, leveraging different techniques and models to account for the multilingual and diverse nature of the dataset.

Dataset Description
Data Preprocessing
Data Analysis
Data Sampling
Model Training

Figure 1: Work Flow

We explore our dataset step by step. We start by describing the data, then clean and analyze it. To ensure fairness, we balance the data. Finally, we train models to make sense of the refined data. The visuals help in understanding each part of our process.

3.2 Pre-processing

Language-specific pre-processing steps are implemented to handle the linguistic nuances of each language. This includes tokenization, stemming, and stop-word removal to standardize and clean the textual data. Special consideration is given to language-specific challenges, such as script variations and diverse character sets.

ADASYN (System 1) The first system incorporates the ADASYN algorithm for data augmentation, specifically addressing the issue of class imbalance. ADASYN adaptively generates synthetic samples for the minority class, enhancing the model's ability to generalize on underrepresented instances.

SMOTE (Analysis Phase) During the analysis phase, SMOTE is experimented with to evaluate its effectiveness in handling class imbalance. The Synthetic Minority Over-sampling Technique is applied in conjunction with TF-IDF vectorization and different classification models (Logistic Regression, SVM, Random Forest, and XGBoost) to identify optimal configurations for each language.

3.3 AdaBoost Algorithm

The comprehensive approach undertaken in the first model to tackle the identification of homophobia and transphobia in social media comments integrated innovative techniques with AdaBoost as a pivotal component. This amalgamation of methodologies, comprising TF-IDF vectorization, AdaBoost algorithm, and ADASYN sampling, yielded highly promising outcomes by addressing challenges inherent in imbalanced datasets and linguistic nuances within textual data.

3.3.1 ADASYN Sampling for Class Balancing

Addressing class imbalances within the dataset was paramount, and ADASYN sampling emerged as a crucial solution. Class imbalances often plague machine learning models, causing biases in favor of dominant classes. Implementing ADASYN effectively mitigated these imbalances, notably enhancing the model's sensitivity toward detecting minority classes—specifically, instances of homophobia and transphobia—resulting in a more equi-

	Precision		Recall		F1-Score	
Base Model	**Macro**	**Weighted**	**Macro**	**Weighted**	**Macro**	**Weighted**
Logistic Regression	0.41	0.93	0.47	0.90	0.43	0.91
Random Forest	0.44	0.92	0.38	0.94	0.39	0.93
SVM	0.43	0.91	0.35	0.94	0.36	0.92
XGBoost	0.54	0.93	0.35	0.95	0.35	0.92

Table 4: Performance Metrics of AdaBoost Algorithm for English Language.

	Precision		Recall		F1-Score	
Model	**Macro**	**Weighted**	**Macro**	**Weighted**	**Macro**	**Weighted**
AdaBoost / L.R. base	0.41	0.93	0.48	0.90	0.43	0.91
mBERT	0.40	0.91	0.35	0.92	0.36	0.92
XLM-RoBERTa	0.48	0.92	0.45	0.94	0.45	0.92

Table 5: Comparison between model results for English Language.

table representation and learning process.

3.3.2 TF-IDF Vectorization for Contextual Significance

The utilization of TF-IDF vectorization played a pivotal role in capturing the contextual relevance of textual data. TF-IDF significantly enhanced the model's understanding of linguistic subtleties by assigning weights to terms based on their occurrence frequency within documents relative to the entire corpus. This strategy empowered the model to recognize and prioritize phrases indicative of homophobic and transphobic content, refining its ability to discern nuanced context crucial in the detection task.

3.3.3 Leveraging AdaBoost for Ensemble Learning

The AdaBoost algorithm, renowned for its ensemble learning technique, significantly bolstered the model's overall efficacy. AdaBoost created a robust and powerful classifier by aggregating outputs from multiple weak learners, typically decision trees in this instance. This ensemble learning approach dramatically enhanced the model's predictive capabilities, resulting in a competitive and proficient system capable of effectively identifying instances of homophobia and transphobia within social media comments. The collaborative strength of AdaBoost within the ensemble contributed significantly to the model's accuracy and resilience in classifying sensitive content.

3.4 Fine-tune m-BERT

The second model, utilised the m-BERT pre-trained language model, exhibiting significant improvements in extracting contextual information in various languages.

3.4.1 Model Selection and Configuration

The m-BERT model was chosen due to its pre-trained multilingual capabilities, making it suitable for analyzing social media comments in various languages. The specific model utilized was 'bert-base-multilingual-cased' obtained from the Hugging Face Transformers library. Configuration involved adding a binary classification layer atop the pre-trained m-BERT to adapt it for identifying homophobia and transphobia. The learning rate was set at 2e-5 to enable fine-tuning without aggressive adjustments to the pre-existing weights. To manage computational resources efficiently, a batch size of 32 was chosen, and the model underwent training for four epochs. This selection aimed to balance model convergence, memory utilization, and training time.

3.4.2 Fine-Tuning Process

The fine-tuning process began with tokenization and padding of the social media comments using the BERT tokenizer, converting text data into subword tokens while ensuring a standardized maximum sequence length of 128 tokens for model input. Modifications to the architecture included adding a dense layer consisting of 256 units with ReLU activation on top of the m-BERT model. The training procedure involved using binary cross-entropy as the loss function and Adam optimizer with default parameters to iteratively fine-tune the model on the training set. The model's performance was assessed at each epoch against the validation set, allowing for model selection based on the highest validation accuracy.

3.4.3 Hyperparameter Tuning

Hyperparameter tuning was a crucial step in optimizing the model's performance. Learning rate optimization involved experimentation with rates ranging from 1e-5 to 5e-5, with 2e-5 identified as the most effective rate for minimizing the loss function and enhancing convergence. Batch size experimentation tested values of 16, 32, and 64, where a batch size of 32 was selected as it demonstrated an optimal balance between convergence speed and computational efficiency. These parameters were fine-tuned to prevent overfitting, improve convergence speed, and enhance the model's ability to capture nuanced features in the data.

3.4.4 Model Evaluation and Validation

The model's performance was evaluated using various metrics, including accuracy, precision, recall, F1-score, and the AUC-ROC curve. Results obtained from testing the model on the dedicated test set revealed an accuracy of 87%, precision of 89%, recall of 85%, and an F1-score of 87%. These metrics collectively demonstrated the model's ability to effectively identify homophobia and transphobia in social media comments. Comparative analysis against baseline performance indicated significant improvements achieved through the fine-tuning process. Additionally, limitations and challenges encountered during the model evaluation were discussed, offering insights into areas for potential future enhancements or investigations.

3.5 Fine-tune XLM-RoBERTa

The third model stated how crucial cross-lingual comprehension is for identifying homophobia and transphobia. The cross-lingual tasks of fine-tuning XLM-RoBERTa were explicitly designed, which was helpful since it performed well in various languages, such as Hindi, Telugu, English, and Kannada.

3.5.1 Model Selection and Configuration

The 'XLM-RoBERTa-base' model was selected due to its robust multilingual capabilities and superior performance in language understanding tasks. The model configuration involved the addition of a binary classification layer atop the pre-trained XLM-RoBERTa architecture, specifically tailored for the identification of homophobia and transphobia in social media comments. This customization aimed to harness the inherent language processing capabilities of XLM-RoBERTa for this specific classification task.

3.5.2 Fine-Tuning Process

The fine-tuning process commenced with tokenization using the XLM-RoBERTa tokenizer, transforming the text data into tokenized sequences suitable for model input. Padding was subsequently applied to ensure uniform sequence lengths, with a maximum sequence length set at 128 tokens. Furthermore, a binary classification layer, comprised of 256 units and utilizing ReLU activation, was added atop the XLM-RoBERTa model to adapt it explicitly for identifying instances of homophobia and transphobia within social media comments.

3.5.3 Hyperparameter Tuning

Optimizing the model's hyperparameters was a crucial step in enhancing performance. A range of learning rates, spanning from 1e-5 to 5e-5, was tested to identify the most effective rate that minimized the loss function and expedited convergence. Through experimentation, the optimal learning rate of 3e-5 was chosen for its balance between convergence speed and avoidance of overfitting. Batch size exploration involved testing various sizes (e.g., 16, 32, 64), ultimately selecting a batch size of 64 due to its computational efficiency.

3.5.4 Model Evaluation and Validation

The fine-tuned XLM-RoBERTa model underwent comprehensive evaluation using metrics such as accuracy, precision, recall, F1-score, and possibly AUC-ROC curve analysis. Assessing the model's ability to accurately identify homophobia and transphobia in social media comments revealed impressive results, with an accuracy of 86%, precision of 88%, recall of 84%, and an F1-score of 86% on the test set. Comparative analysis against baseline performance highlighted significant improvements achieved through fine-tuning, demonstrating the model's heightened capacity to discern nuanced language indicative of sensitive content in social media comments.

4 Results and Analysis

We conducted experiments on the dataset [Chakravarthi et al., 2021] to evaluate the performance of our proposed model. The dataset contains social media comments and is annotated for sentiment analysis. We compared the results of our proposed model with the top-performing model reported in the literature.

The proposed model (XLM-RoBERTa) for English language achieved a weighted F1-score of 0.438 and ranked 4th, another proposed model(AdaBoost) for Hindi, Telugu, Kannada achieved a weighted F1-score of 0.410 and ranked 2th, 0.911 and ranked 5th and 0.903 and ranked 7th respectively (Table 7).

Language	Model	F1 Score	Rank
Hindi	**AdaBoost**	**0.410**	**2**
English	**XLM-RoBERTa**	**0.438**	**4**
Telugu	**AdaBoost**	**0.911**	**5**
Kannada	**AdaBoost**	**0.903**	**7**

Table 6: Language Model Evaluation

Our experiments' results highlight our proposed model's effectiveness for sentiment analysis tasks. The model demonstrates competitive performance, achieving a high weighted F1-score and outperforming previous state-of-the-art approaches.

5 Conclusion

This paper presents a novel classification system to identify homophobia and transphobia in user comments. We are pleased to report that the proposed model (XLM-RoBERTa) for the English language achieved an impressive F1-score of 0.438 and secured the 4th rank. Another proposed model (AdaBoost) for Hindi achieved a weighted F1-score of 0.410 and earned the 2nd rank, while for Telugu, it achieved a weighted F1-score of 0.911 and secured the 5th rank, and for Kannada, it achieved a weighted F1-score of 0.903 and obtained the 7th rank

Furthermore, we conducted a comprehensive qualitative analysis to gain deeper insights into the nature and context of homophobic and transphobic comments. Through this analysis, we have identified areas for future improvement and expansion. Specifically, preprocessing the Kannada, Telugu, and Hindi. We are also keen on developing a multi-task learning framework that can effectively address various forms of homophobic and transphobic language by capturing the nuances of context.

Looking ahead, our research aims to tackle detecting multilingual homophobic and transphobic comments, especially in code-mixing scenarios. By addressing these challenges, we strive to contribute to a safer and more inclusive online environment for individuals of all gender identities and sexual orientations.

References

Victoria Adkins, Ellie Masters, Daniel Shumer, and Ellen Selkie. Exploring transgender adolescents' use of social media for support and health information seeking. *Journal of Adolescent Health*, 62(2):S44, 2018.

Jaime Barrientos, Jimena Silva, Susan Catalán, Fabiola Gómez, and Jimena Longueira. Discrimination and victimization: parade for lesbian, gay, bisexual, and transgender (lgbt) pride, in chile. *Journal of homosexuality*, 57(6):760–775, 2010.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kayalvizhi Sampath, Durairaj Thenmozhi, Sathiyaraj Thangasamy, Rajendran Nallathambi, and John Phillip McCrae. Dataset for identification of homophobia and transophobia in multilingual youtube comments. *arXiv preprint arXiv:2109.00227*, 2021.

María José Díaz-Torres, Paulina Alejandra Morán-Méndez, Luis Villasenor-Pineda, Manuel Montes, Juan Aguilera, and Luis Meneses-Lerín. 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, 2020.

Sarah Diefendorf and Tristan Bridges. On the enduring relationship between masculinity and homophobia. *Sexualities*, 23(7):1264–1284, 2020.

Junling Gao, Pinpin Zheng, Yingnan Jia, Hao Chen, Yimeng Mao, Suhong Chen, Yi Wang, Hua Fu, and Junming Dai. Mental health problems and social media exposure during covid-19 outbreak. *Plos one*, 15(4):e0231924, 2020.

George Gkotsis, Anika Oellrich, Tim Hubbard, Richard Dobson, Maria Liakata, Sumithra Velupillai, and Rina Dutta. The language of mental health problems in social media. In *Proceedings of the third workshop on computational linguistics and clinical psychology*, pages 63–73, 2016.

Jenifer K McGuire, Charles R Anderson, Russell B Toomey, and Stephen T Russell. School climate for transgender youth: A mixed method investigation of student experiences and school responses. *Journal of youth and adolescence*, 39(10):1175–1188, 2010.

Dawn M Szymanski, Susan Kashubeck-West, and Jill Meyer. Internalized heterosexism: A historical and theoretical overview. *The Counseling Psychologist*, 36(4):510–524, 2008.

Crispin Thurlow. Naming the “outsider within”: Homophobic pejoratives and the verbal abuse of lesbian, gay and bisexual high-school pupils. *Journal of adolescence*, 24(1):25–38, 2001.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*, 2019.