KEC-AI-NLP@LT-EDI-2024:Homophobia and Transphobia Detection in Social Media Comments using Machine Learning

Kogilavani Shanmugavadivel[1], **Malliga Subramanian**[1], **Shri Durga R**[1],
Srigha S[1], **Samyuktha K**[1], **Nithika K**[1]
[1]Department of AI, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{shridurgar.21aim, srighas.21aim}@kongu.edu
{samyukthak.21aim,nithikak.21aim}@kongu.edu

Abstract

Our work addresses the growing concern of abusive comments in online platforms, particularly focusing on the identification of Homophobia and Transphobia in social media comments. The goal is to categorize comments into three classes: Homophobia, Transphobia, and non-anti LGBT+ comments. Utilizing machine learning techniques and a deep learning model, our work involves training on a English dataset with a designated training set and testing on a validation set. This approach aims to contribute to the understanding and detection of Homophobia and Transphobia within the realm of social media interactions. Our team participated in the shared task organized by LT-EDI@EACL 2024[1] and secured seventh rank in the task of Homophobia/Transphobia Detection in social media comments in Tamil with a macro- f1 score of 0.315. Also, our run was submitted for the English language and secured eighth rank with a macro-F1 score of 0.369. The run submitted for Malayalam language securing fourth rank with a macro- F1 score of 0.883 using the Random Forest model.

1 Introduction

In the contemporary digital landscape, social media platforms serve as pivotal mediums for communication, education, and information sharing. Among these platforms, YouTube stands out as a prominent social networking and video-sharing hub, enabling users to create accounts, share videos, and interact through comments. However, the prevalence of abusive comments, particularly targeting transgender and homosexual individuals, poses a significant challenge to the well-being of platform users. The escalating use of online communication has raised concerns about the dissemination of slander, hate speech, and cyberbully, with negative consequences for individuals and societal harmony. Slander, characterized by false spoken statements that harm individuals or groups, is increasingly acknowledged for its detrimental impact Olweus and Limber (2018). Such negative comments not only inflict psychological harm but also contribute to the proliferation of animosity, division, and discontent in online spacesMishna et al. (2009). Major social media platforms like YouTube, Facebook, Instagram, and Twitter have responded by implementing policies and protocols to address and mitigate hateful content. Our study aims to scrutinize and identify offensive comments within an English dataset, treating the detection of abusive comments as a text classification problem. Focused on machine learning and deep learning methodologies, our research excludes the use of transfer learning models and does not involve the integration of machine learning and deep learning approaches. The objective is to train and compare various models to determine the optimal approach for identifying hate comments in English.

2 Literature Review

Research in the field of abusive language detection spans various approaches and methodologies, as evident in several notable papers. Mubarak et al. (2017) emphasize the challenges faced in Arabic abusive language detection, including dialects and informal language. Mishra et al. (2019) introduce a novel approach using Graph Convolutional Networks (GCNs) to capture syntactic and semantic dependencies for effective abusive language identification.

Addressing gender bias in abusive language detection, Park et al. (2018) propose a method incorporating gender information into the training process, showcasing its effectiveness in reducing bias while maintaining overall performance. Ibrohim

[1] https://codalab.lisn.upsaclay.fr/competitions/16056

Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 200–205
March 21, 2024 ©2024 Association for Computational Linguistics

and Budi (2019) focus on multi-label hate speech detection in Indonesian Twitter, analyzing various approaches, including feature-based, deep learning, and ensemble methods.

Narang and Brew (2020) present an approach utilizing syntactic dependency graphs for abusive language detection, achieving superior performance compared to baseline models. Caselli et al. (2021) introduce HateBERT, a retraining approach for BERT tailored for English abusive language detection, demonstrating its superiority in precision, recall, and F1-score.

Davidson et al. (2019) investigate racial bias in hate speech datasets, highlighting potential biases in annotation processes and emphasizing the need for fair evaluations. Koufakou et al. (2020) introduce HurtBERT, combining BERT with lexical features for enhanced abusive language detection performance.

Corazza et al. (2020) propose a zero-shot abusive language detection using emoji-based masked language models, demonstrating competitive performance. Chakravarthi (2020) contribute HopeEDI, a multilingual dataset for hope speech detection, aiming to facilitate research on positive discourse in social media.

Overall, these works offer diverse insights and methodologies, advancing the understanding and detection of abusive language in various linguistic and societal contexts.

3 Dataset Description

The goal of this shared task on homophobia and transphobia comment detection is to detect and reduce abusive comments on social media that target homosexual and trans-gender individuals. The dataset used here is shared by the shared task Chakravarthi et al. (2023). The primary goal of this project is to develop methods for detecting and classifying instances of hate speech in English language. The Homophobia and Transphobia Comment Detection data set is made up of English comments retrieved from the YouTube comments area Kumaresan et al. (2023). The data set consists of a comment and its related label from one of the three labels: Non-anti-LGBT+ content, Homophobia, Transphobia. SMOTE, which stands for Synthetic Minority Over-sampling data augmentation Technique, is a widely used technique in the field of machine learning specifically in the context of handling imbalanced datasets.Imbalanced datasets occur when the classes have significantly different numbers of instances, leading to a bias in the model's performance towards the majority class.

3.1 English Data

The Train, Test, and Development data sets each comprise 3,164, 792, 991 comments which is summarized in Table 1. The text in English is followed by the appropriate label for each comment in the training data. As Table 2 suggests, the Transphobia label exhibits a significant scarcity, leading to a pronounced class imbalance. Due to the limited availability of test or development data examples for the Transphobia label, the classification task becomes particularly challenging, focusing predominantly on the other two labels.

Table 1: Data-set Description

Data-set	No. of Comments
Train	3,164
Validation	792
Test	991

Table 2: Class Description

Class	Train	Dev	Test
Non-anti-LGBT+	2,978	748	931
Homophobia	179	43	55
Transphobia	7	2	4

4 Methodology

Machine learning and deep learning models cannot access raw texts. Feature extraction is required to train classification models. The TF-IDF representation is utilized in ML techniques to extract features. Figure 1 gives the detailed workflow of our proposed model. We use three ways to analyze the results and create the best model possible: Machine Learning, Deep Learning.

4.1 Machine Learning Models

Machine learning has come a long way in recent years, changing the way people understand important applications such as image recognition, data mining, and natural language processing(NLP). This section outlines the machine learning models utilized in the present study for text classification. We used several different kinds of machine learning algorithms such as Decision tree, Random Forest, GaussianNB, XGBoost, AdaBoost, KNN,

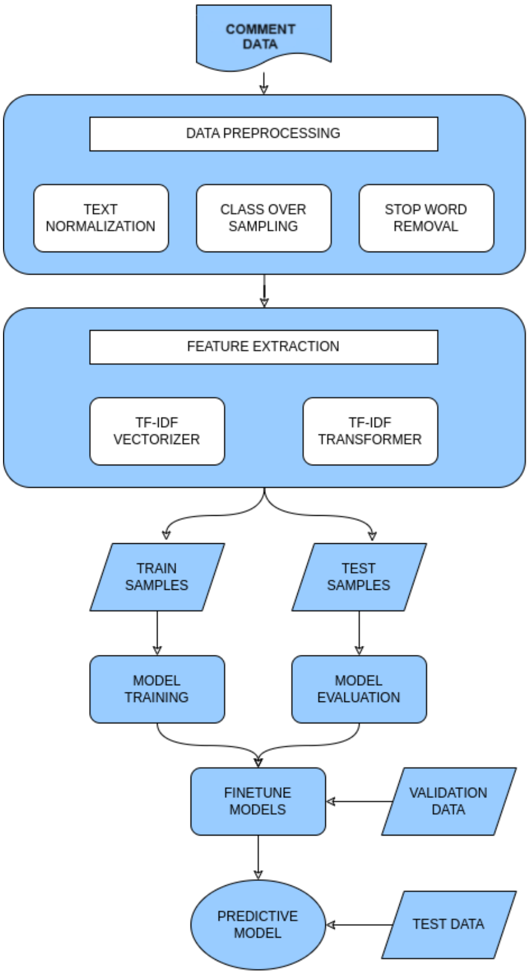


Figure 1: Proposed System Workflow

Linear Regression, Multinomial NB, Support Vector Machine, MLP Classifier, Gradient Boost, and Ensemble models.

4.2 Feature Extraction

The TF-IDF Vectorizer with Character N-grams is a feature extraction technique widely employed in Natural Language Processing (NLP) for the effective representation of textual data in machine learning models. Operating at the character level, this vectorizer analyzes individual characters rather than complete words, allowing it to capture sequential patterns within the text. The inclusion of character n-grams, specified here with lengths ranging from 1 to 3, proves particularly advantageous in tasks that demand consideration of word morphology and character-level nuances, such as sentiment analysis or language-specific challenges. The TF-IDF weighting scheme assigns significance weights to these character n-grams based on their occurrence within individual documents and across the entire dataset. This method not only enhances the representation of textual information but also facilitates the identification of key character patterns. The limitation of the feature space to the top most influential n-grams ensures a focused and meaningful representation, contributing to the efficiency of subsequent machine learning algorithms.

4.3 Deep Learning Model

In the realm of homophobia and transphobia detection within English YouTube comments, this study highlights the efficacy of deep learning models, specifically Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. While CNN excels in capturing localized patterns, LSTM proves valuable in handling long-term dependencies in sequential data, making it suitable for comment analysis.

The pre-processed comments undergo LSTM model training and evaluation, where the LSTM network, belonging to the family of recurrent neural networks (RNNs), excels in capturing long-term dependencies within the sequential nature of text data. By considering the temporal information of comments, the LSTM model effectively captures the context and dependencies that exist between words and phrases. This nuanced understanding contributes to the model's ability to discern patterns and relationships within comment sequences, providing a robust foundation for homophobia and transphobia detection in English YouTube comments.

5 Performance Evaluation

After submitting the run using the Random Forest model, it proved beneficial for various languages. Analyzing the results in Table 3, which provides the macro-average of precision, recall, and F1-score for the various models used. Random Forest surpassed both deep learning and other machine learning models in precision, recall, and F1 score. Leveraging an ensemble of decision trees and feature importance estimation, this model effectively captured complex patterns within the dataset

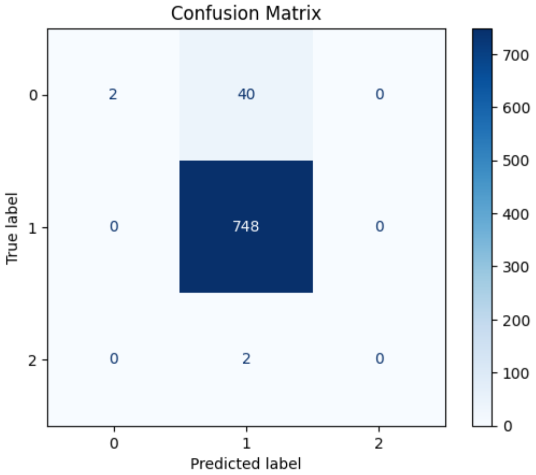
The Random Forest model excelled in handling high-dimensional data, managing noisy and missing values, and mitigating overfitting concerns through feature subsampling and bootstrap aggregating. Notably, the dataset's class distribution was not uniform, with two crucial classes having very few instances. Despite this challenge, the Random Forest model demonstrated exceptional performance.

The contrasting deep learning model, reliant on significant computational resources and extensive parameter tuning, fell short, resulting in comparatively lower accuracy and F1 score. In Figure 2, the presented confusion matrix provides a comprehensive overview of the performance of the Random Forest model when applied to the Malayalam dataset. This emphasizes the importance of selecting an appropriate modeling technique tailored to the dataset's characteristics, leading to improved predictive performance.

6 Conclusion

The study concentrates on detecting homophobic and transphobic comments in YouTube discussions, comparing the performance of various models in this task. Strikingly, Deep Learning models did not demonstrate superior results when trained and evaluated on English data. Instead, Machine Learning models outperformed Deep Learning in effectiveness. It's crucial to note that our study did not make use of contextualized embeddings like BERT or GPT, which have shown potential in enhancing language model performance.

Figure 2: Confusion Matrix of Random Forest Classifier Model



Acknowledging this limitation, we propose that future research should explore the implementation of contextualized embeddings using deep learning techniques, such as BERT or GPT. The absence of these advanced embeddings may have limited the effectiveness of the models used in our study. Incorporating such embeddings holds promise for significantly improving the detection of homophobic and transphobic comments in YouTube discussions. Additionally, we did not explore transfer learning with other models in our current stage. Still, we emphasize the possibility of integrating these models in our future work, indicating a pathway for ongoing exploration and enhancement in identifying such comments on YouTube.

References

Tommaso Caselli, Arjan Schelhaas, Marieke Weultjes, Folkert Leistra, Hylke van der Veen, Gerben Timmerman, and Malvina Nissim. 2021. DALC: the Dutch abusive language corpus. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 54–66.

Bharathi Raja Chakravarthi. 2020. HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53.

Table 3: English Data Evaluation Metrics

Model	Precision	Recall	F1-score
Multilayer perceptron	0.46	0.43	0.44
K-Nearest Neighbour	0.43	0.39	0.40
Xtreme Gradient Boost	0.45	0.35	0.35
Decision Tree	0.37	0.37	0.37
Logistic Regression	0.65	0.34	0.34
Random Forest	0.65	0.34	0.34
Support Vector Classifier	0.57	0.34	0.37
Multinomial Naive Bayes	0.31	0.33	0.32
Gradient Boost Classifier	0.41	0.37	0.39
Ensemble	0.65	0.34	0.34
Adaboost Classifier	0.34	0.34	0.34
CNN-LSTM	0.50	0.37	0.39

Bharathi Raja Chakravarthi. 2023. Detection of homophobia and transphobia in YouTube comments. *International Journal of Data Science and Analytics*, pages 1–20.

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 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 and Inclusion*, Malta. European Chapter of the 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 English, Spanish, Hindi, Tamil, and Malayalam". In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–22.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the third workshop on abusive language online*, pages 46–57.

Simrat Kaur, Sarbjeet Singh, and Sakshi Kaushal. 2021. Abusive content detection in online user-generated data: a survey. *Procedia Computer Science*, 189:274–281.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, Viviana Patti, et al. 2020. HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the fourth workshop on online abuse and harms*, pages 34–43. Association for Computational Linguistics.

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*, page 100041.

Abulimiti Maimaitituoheti. 2022. ABLIMET@ LT-EDI-ACL2022: a RoBERTa based approach for homophobia/transphobia detection in social media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 155–160.

Faye Mishna, Alan McLuckie, and Michael Saini. 2009. Real-world dangers in an online reality: A qualitative study examining online relationships and cyber abuse. *Social Work Research*, 33(2):107–118.

Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Abusive language detection with graph convolutional networks. *arXiv preprint arXiv:1904.04073*.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on Arabic social media. In *Proceedings of the first workshop on abusive language online*, pages 52–56.

Kanika Narang and Chris Brew. 2020. Abusive language detection using syntactic dependency graphs. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 44–53.

Dan Olweus and Susan P Limber. 2018. Some problems with cyberbullying research. *Current opinion in psychology*, 19:139–143.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*.

Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. 2011. Detecting and tracking political abuse in social media. In *Proceedings of the International AAAI Conference on Web and social media*, volume 5, pages 297–304.

Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumar Kumaresan. 2022. Overview of Abusive Comment Detection in Tamil-ACL 2022. *DravidianLangTech*, 2022:292.