SSNCSE_NLP@LT-EDI-ACL2022: Homophobia/Transphobia Detection in Multiple Languages using SVM Classifiers and BERT-based Transformers

Krithika Swaminathan
SSN College of Engineering
krithika2010039@ssn.edu.in

Hrishik Sampath
SSN College of Engineering
hrishik2010483@ssn.edu.in

Gayathri G L
SSN College of Engineering
gayathri2010090@ssn.edu.inn

B. Bharathi
SSN College of Engineering
bharathib@ssn.edu.in

Abstract

Over the years, there has been a slow but steady change in the attitude of society towards different kinds of sexuality. However, on social media platforms, where people have the license to be anonymous, toxic comments targeted at homosexuals, transgenders and the LGBTQ+ community are not uncommon. Detection of homophobic comments on social media can be useful in making the internet a safer place for everyone. For this task, we used a combination of word embeddings and SVM Classifiers as well as some BERT-based transformers. We achieved a weighted F1-score of 0.93 on the English dataset, 0.75 on the Tamil dataset and 0.87 on the Tamil-English Code-Mixed dataset.

1 Introduction

Human beings have constantly tried to create an identity for themselves, and with the world becoming increasingly progressive, they have more freedom of choice in many spheres of life, including gender expressions and sexuality.(Cederved et al., 2021) However, the understanding of these concepts continues to gradually evolve, and despite various major social advancements in the last few years, LGBTQIA+ people face discrimination on the grounds of sexual orientation and gender identity.

Although social media has provided this minority with a platform to express themselves by sharing their experiences and build a strong, healthy community, there has been an increasing amount of general toxicity on the internet (Craig and McInroy, 2014). There has also been a spread of transphobic and homophobic comments through these online forums, due to the easy access to anonymity they provide, which ensures that these violators are never held accountable(McInroy and Craig, 2015)(Gámez-Guadix and Incera, 2021).

The need for the detection and filtering of such acerbic content in user-created online content is thus at an all-time high. However, the manual detection and flagging of certain words might be time-consuming and ineffective in the long run. The tendency of Tamil speakers to use code-mixed transliterated text also poses a challenge to the task.

In this paper, we examine various approaches for the classification of Tamil code-mixed comments into three categories, namely, Homophobic, Transphobic and Non-anti-LGBT+ content as a part of the shared task Homophobia/Transphobia Detection @ LT-EDI-ACL2022 (Chakravarthi et al., 2022a).

After tackling the data imbalance using sampling techniques, feature extraction using count vectorizer and tf-idf was done along with various classifiers. Another approach involved the usage of transformer models to classify the text. The same has also been analysed for English and Tamil datasets.

The remainder of the paper is organized as follows. Section 2 discusses related works according to this task. Section 3 analyses the given datasets. Section 4 outlines the methodology followed for the task. The results are presented in Section 5 and finally, a conclusion is delivered.

2 Related Work

The first formal defense of homosexuality was published in 1908 (Edsall, 1908). The 20th century witnessed many ups and downs in the progress of social acceptance of sexual minorities. Various studies on the existence of different sexualities have been conducted such as (Ventriglio and Bhugra, 2019), (Francis et al., 2019), (Trinh, 2022) and (Kiesling, 2019), and it has been observed that there has been a positive shift in the attitude of the general public towards homosexuality (Cheng et al., 2016) (Mathews et al., 1986). More recently, the LGBTQ+ movement has picked up and has gained many followers through social media. Several people have worked on the task of using machine learning to identify and filter

Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 239 - 244
May 27, 2022 ©2022 Association for Computational Linguistics

out hurtful comments, thus aiding in the battle against homophobic/transphobic sentiments. Some of the early works in this field include (Mandl et al., 2020) and (Díaz-Torres et al., 2020), in which offensive language is identified in multiple Indian languages as well as some foreign languages. In (Pereira, 2018), homophobia was predicted in Portuguese tweets using supervised machine learning and sentiment analysis techniques. A wide range of techniques was utilised in this study, some of which include Naive Bayes, Random Forest and Support Vector Machines. The models were combined using voting and stacking, with the best results being obtained through voting using 10 models. (Chakravarthi et al., 2021) presents an expert-labelled dataset and various machine learning models for the identification and classification of Homophobia and Transphobia in multilingual YouTube Comments. In (Chakravarthi et al., 2022b), sentiment analysis and offensive language detection were performed for Dravidian languages in code-mixed text, which are super-sets of the Homophobia/Transphobia detection task. In this paper, an experimentation of a number of machine learning algorithms such as SVM, MNB, KNN, Decision Tree, Random Forest and some BERT-based transformers, was done.

In our work, we have put forward a comparison of some of the most popular models for this area of research and estimated the top three models for each language in the datasets given for this task.

3 Dataset Analysis and Preprocessing

Category	English	Tamil	Tamil-English
Homophobic	276	723	465
Transphobic	13	233	184
Non-anti-LGBT+ content	4657	3205	5385
Total	4946	4161	6034

Table 1: Data distribution of training dataset

The three datasets given for this Homophobia/Transphobia Detection task are sets of comments from social media platforms, primarily YouTube, with the data given in the languages English, Tamil and Tamil-English code-mixed. The comments in these datasets are classified into one of these three categories - Homophobic, Transphobic and Non-anti-LGBT+ content. Table 1 outlines the data distribution of each training dataset. Most of the comments in these datasets do not extend beyond a single sentence and the average number of sentences in each comment is close to 1.

All three datasets are highly imbalanced with respect to the categorisation classes. Considering this imbalance in the data distribution, it is expected that training a model on these datasets would give rise to a bias in the predictions towards the dominant category class in each dataset. Figure 1 illustrates the highly disproportionate distribution of data in each of the given datasets.

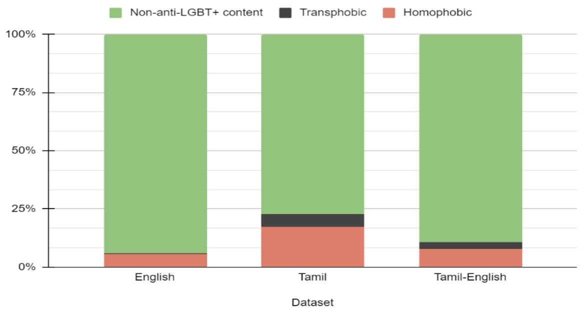

Figure 1: Graphical representation of data distribution

The given raw datasets may contain inconsistencies in their data or may contain unnecessary data. Before feeding the data to the required algorithm, it is therefore important to clean the datasets. This cleansing of the datasets is carried out by removing punctuation, special characters and excess words that semantically contribute nothing to the overall mood of each comment.

4 Methodology

As part of our experimental setup, various classifier models were applied to the processed data after extracting the necessary features from it. For each dataset, three models that worked best for the language under consideration were chosen to predict the classification results for comments collected in that language.

For reference, the models under consideration for the English dataset have been listed in Table 6 and Table 7 along with their performance on the development data. Similarly, the performance of the models for the Tamil dataset has been tabulated in Table 8 and Table 9, and their performance on the Tamil-English dataset has been illustrated in Table 10 and Table 11.

Feature	Classifier	Precision	Recall	F1-score	Accuracy
Count vectorizer	SVM	0.51	0.38	0.40	0.93
Indo Aryan XLM R-Base transformer	SVM	0.53	0.39	0.42	0.93
Average_word_embeddings_glove_6B_300d	SVM	0.54	0.40	0.44	0.94

Table 2: Performance of the proposed approach of English text using dev data

Feature	Classifier	Precision	Recall	F1-score	Accuracy
Count vectorizer	SVM	0.86	0.66	0.73	0.89
TF-IDF	SVM	0.88	0.84	0.86	0.94
Transformer monsoon-nlp/tamillion	-	0.56	0.60	0.58	0.90

Table 3: Performance of the proposed approach of Tamil text using dev data

4.1 Embedding

Embedding is used to encode the meaning of words in a text by transforming them into real-valued vectors. After successful embedding, words with similar meanings are found to be grouped together. For this task, we experimented using some BERT-based sentence transformer models and word embeddings.

4.2 Feature extraction

A feature is a unique property of a text by which it can be measured or quantified. Feature extraction helps to reduce the complexity of dataset on which a model is to be trained. Numeric encoding of the text is done as a part of this process.

4.2.1 Feature extraction using Count vectorizer

The Count Vectorizer is used to tokenize a set of texts by converting the collection of texts to a vector of token counts. The strategies of tokenization, counting and normalization are together called as the n-gram representation.

4.2.2 Feature extraction using TF-IDF

TF-IDF, which stands for term frequency-inverse document frequency, is a method of quantifying a sentence based on the words in it. Each row is vectorized using a technique in which a score is computed for each word to signify their importance in the text. The score for commonly used words is decreased while the score for rare words is increased.

4.3 Models applied

Some models that we experimented on for this task include Classifiers such as SVM, NLP, random forest and K-nearest neighbours, and some simple transformers like LaBSE, tamillion and IndicBERT. These experiments were conducted for English, Tamil and Tamil-English code-mixed data. The best models observed were selected to generate the performance scores for the data sets.

5 Observations

It was found that certain models or combinations of models outperform others for each dataset under scrutiny. The performance results for each chosen model are presented in the tables given below.

This task is evaluated on the macro averages of three performance metrics - Precision, Recall and F1-score. The scores achieved for this Homophobia Detection task are tabulated below in Table 5.

5.1 English dataset

After the required features were extracted, they were trained with different machine learning models. The models were then evaluated using the development data. The performance of the chosen models on the development data of the English dataset is depicted in Table 2.

Our submission secured the 11th rank in Task B, i.e., Homophobia/Transphobia Detection on an English dataset. Our model procured a macro F1-Score of 0.37 and a weighted F-score of 0.93.

5.2 Tamil dataset

After the required features were extracted, they were trained with different machine learning models. The models were then evaluated using the development data. The performance of the chosen models on the development data of the Tamil dataset is depicted in Table 3.

Our submission secured the 9th rank in Task B, i.e., Homophobia/Transphobia Detection on a

Feature	Classifier	Precision	Recall	F1-score	Accuracy
Count vectorizer	SVM	0.71	0.44	0.48	0.90
TF-IDF	SVM	0.67	0.54	0.58	0.89
Transformer setu4993/LaBSE	-	0.70	0.50	0.55	0.90

Table 4: Performance of the proposed approach of Tamil-English text using dev data

Dataset	Accuracy	macro Precision	macro Recall	macro F1-score	Weighted Precision	Weighted Recall	Weighted F1-score	Rank
English	-	0.93	0.48	0.37	0.39	0.91	0.93	11
Tamil	0.77	0.55	0.47	0.50	0.74	0.77	0.75	9
Tamil-English	0.89	0.66	0.43	0.47	0.87	0.89	0.87	9

Table 5: Performance scores for the Homophobia Detection task

Tamil dataset. Our model procured a macro F1-Score of 0.50 and a weighted F-score of 0.75.

5.3 Tamil-English dataset

After the required features were extracted, they were trained with different machine learning models. The models were then evaluated using the development data. The performance of the chosen models on the development data of the Tamil-English code-mixed dataset is depicted in Table 4.

Our submission secured the 9th rank in Task B, i.e., Homophobia/Transphobia Detection on a Tamil-English code-mixed dataset. Our model procured a macro F1-Score of 0.47 and a weighted F-score of 0.87.

5.4 Inferences

It is observed that each of the datasets is not very large and therefore, the number of training samples is limited. Almost all the classifier and transformer models used made highly accurate predictions on the English dataset. For the Tamil and Tamil-English code-mixed datasets, there is a significant variation in the performances of the different models used. It is evident that the SVM and MLP classifier models have similar good accuracy rates after performing some feature extraction, with SVM having a slight edge over MLP. The overall performance of the TF-IDF model is found to be slightly higher than that of the count vectorizer model. For the datasets with Tamil text, sentence transformers pre-trained for multilingual texts performed well. The LaBSE model was found to work particularly well for Tamil text. In summary, the SVM classifier model and the LaBSE transformer model yielded the best results for this classification task.

Conclusion

In this study, we have presented a comparison of different models for the LT-EDI-ACL 2022 shared task on homophobia detection. It was observed that average word embeddings along with the SVM Classifier worked the best for English text and that a combination of the tf-idf vectorizer and the SVM Classifier performed well on Tamil text. A language agnostic model called LaBSE worked best for Tamil-English code-mixed text. These results can further be improved by using suitable embeddings for each model and employing better preprocessing techniques.

References

Catarina Cederved, Stinne Glasdam, and Sigrid Stjernswärd. 2021. A clash of sexual gender norms and understandings: A qualitative study of homosexual, bisexual, transgender, and queer adolescents' experiences in junior high schools. *Journal of Adolescent Research*, page 07435584211043290.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John Phillip McCrae, Paul Buitaleer, Prasanna Kumar Kumaresan, and Rahul Ponnusamy. 2022a. Findings of the shared task on Homophobia Transphobia Detection in Social Media Comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P McCrae. 2022b. Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. *Language Resources and Evaluation*, pages 1–42.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan,

S.No.	Feature Extraction	Classifier	Precision	Recall	F1-score	Accuracy
1	Count Vectorizer	SVM	0.51	0.38	0.40	0.93
2	Count Vectorizer	K nearest neighbour	0.53	0.36	0.36	0.92
3	Count Vectorizer	MLP Classifier	0.52	0.40	0.43	0.92
4	TF-IDF	SVM	0.48	0.38	0.40	0.92
5	TF-IDF	K nearest neighbour	0.44	0.37	0.38	0.92
6	TF-IDF	MLP Classifier	0.48	0.34	0.33	0.92

Table 6: Performance of the selected classifier models on English text using dev data

S.No.	Pre-trained model	Precision	Recall	F1-score	Accuracy
1	distilbert-base-uncased-finetuned-sst-2-english	0.43	0.44	0.43	0.90
2	Indo-Aryan-XLM-R-Base	0.53	0.39	0.42	0.93
3	average_word_embeddings_glove.6B.300d	0.48	0.34	0.33	0.94

Table 7: Performance of the selected transformer models on English text using dev data

S.No.	Feature Extraction	Classifier	Precision	Recall	F1-score	Accuracy
1	Count Vectorizer	SVM	0.86	0.66	0.73	0.89
2	Count Vectorizer	K nearest neighbour	0.58	0.53	0.54	0.75
3	Count Vectorizer	MLP Classifier	0.89	0.78	0.82	0.93
4	TF-IDF	SVM	0.88	0.84	0.86	0.94
5	TF-IDF	K nearest neighbour	0.61	0.64	0.59	0.76
6	TF-IDF	MLP Classifier	0.80	0.90	0.73	0.90

Table 8: Performance of the selected classifier models on Tamil text using dev data

S.No.	Pre-trained model	Precision	Recall	F1-score	Accuracy
1	bert-base-multilingual-uncased	0.84	0.52	0.54	0.84
2	setu4993/LaBSE	0.86	0.88	0.87	0.94
3	monsoon-nlp/tamillion	0.56	0.60	0.58	0.90

Table 9: Performance of the selected transformer models on Tamil text using dev data

Kayalvizhi Sampath, Durairaj Thenmozhi, Sathiyaraj Thangasamy, Rajendran Nallathambi, and John Phillip McCrae. 2021. Dataset for identification of homophobia and transophobia in multilingual youtube comments. *arXiv preprint arXiv:2109.00227*.

Yen-hsin Alice Cheng, Fen-Chieh Felice Wu, and Amy Adamczyk. 2016. Changing attitudes toward homosexuality in taiwan, 1995–2012. *Chinese Sociological Review*, 48(4):317–345.

Shelley L Craig and Lauren McInroy. 2014. You can form a part of yourself online: The influence of new media on identity development and coming out for lgbtq youth. *Journal of Gay & Lesbian Mental Health*, 18(1):95–109.

María José Díaz-Torres, Paulina Alejandra Morán-Méndez, Luis Villasenor-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.

Nicholas C. Edsall. 1908. *Toward Stonewall: Homosexuality and Society in the Modern Western World.*

Dennis A Francis, Anthony Brown, John McAllister,

S.No.	Feature Extraction	Classifier	Precision	Recall	F1-score	Accuracy
1	Count Vectorizer	SVM	0.71	0.44	0.48	0.90
2	Count Vectorizer	K nearest neighbour	0.55	0.41	0.44	0.88
3	Count Vectorizer	MLP Classifier	0.69	0.45	0.50	0.89
4	TF-IDF	SVM	0.67	0.54	0.58	0.89
5	TF-IDF	K nearest neighbour	0.68	0.43	0.47	0.86
6	TF-IDF	MLP Classifier	0.73	0.39	0.41	0.89

Table 10: Performance of the selected classifier models on Tamil-English text using dev data

S.No.	Pre-trained model	Precision	Recall	F1-score	Accuracy
1	bert-base-multilingual-uncased	0.38	0.37	0.37	0.88
2	setu4993/LaBSE	0.70	0.50	0.55	0.90
3	monsoon-nlp/tamillion	0.35	0.34	0.33	0.89

Table 11: Performance of the selected transformer models on Tamil-English text using dev data

Sethunya T Mosime, Glodean TQ Thani, Finn Reygan, Bethusile Dlamini, Lineo Nogela, and Marguerite Muller. 2019. A five country study of gender and sexuality diversity and schooling in southern africa. *Africa Education Review*, 16(1):19–39.

Manuel Gámez-Guadix and Daniel Incera. 2021. Homophobia is online: Sexual victimization and risks on the internet and mental health among bisexual, homosexual, pansexual, asexual, and queer adolescents. *Computers in human behavior*, 119:106728.

Scott F Kiesling. 2019. *Language, Gender, and Sexuality: An Introduction*. Routledge.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Forum for Information Retrieval Evaluation*, pages 29–32.

William C Mathews, Mary W Booth, John D Turner, and Lois Kessler. 1986. Physicians' attitudes toward homosexuality–survey of a california county medical society. *Western Journal of Medicine*, 144(1):106.

Lauren B McInroy and Shelley L Craig. 2015. Transgender representation in offline and online media: Lgbtq youth perspectives. *Journal of Human Behavior in the Social Environment*, 25(6):606–617.

Vinicius Gomes Pereira. 2018. *Using supervised machine learning and sentiment analysis techniques to predict homophobia in portuguese tweets*. Ph.D. thesis.

Ethan Trinh. 2022. Supporting queer slife youth: Initial queer considerations. In *English and Students with Limited or Interrupted Formal Education*, pages 209–225. Springer.

Antonio Ventriglio and Dinesh Bhugra. 2019. Sexuality in the 21st century: Sexual fluidity. *East Asian Archives of Psychiatry*, 29(1):30–34.