Quartet@LT-EDI 2024: Support Vector Machine Based Approach For Homophobia/Transphobia Detection In Social Media Comments

Shaun Allan H
Sri Sivasubramaniya Nadar College of Engineering
shaunallan2210716@ssn.edu.in

Samyuktaa Sivakumar
Sri Sivasubramaniya Nadar College of Enginerring
samyuktaa2210189@ssn.edu.in

Rohan R
Sri Sivasubramaniya Nadar College of Enginerring
rohan2210124@ssn.edu.in

Nikilesh Jayaguptha
Sri Sivasubramaniya Nadar College of Engineering
nikilesh2210219@ssn.edu.in

Durairaj Thenmozhi
Sri Sivasubramaniya Nadar College of Engineering
thenid@ssn.edu.in

Abstract

Homophobia and transphobia are terms which are used to describe the fear or hatred towards people who are attracted to the same sex or people whose psychological gender differs from his biological sex. People use social media to exert this behaviour. The increased amount of abusive content negatively affects people in a lot of ways. It makes the environment toxic and unpleasant to LGBTQ+ people. The paper talks about the classification model for classifying the contents into 3 categories which are homophobic, transphobic and non-homophobic/transphobic. We used many traditional models like Support Vector Machine, Random Classifier, Logistic Regression and K-Nearest Neighbour to achieve this. The macro average F1 scores for Malayalam, Telugu, English, Marathi, Kannada, Tamil, Gujarati, Hindi are 0.88, 0.94, 0.96, 0.78, 0.93, 0.77, 0.94, 0.47 and the rank for these languages are 5, 6, 9, 6, 8, 6, 6, 4.

1 Introduction

Social media platforms of the current century have evolved into a way people communicate with each other. Social media is about having conversations, building communities and connecting with the audience. It has become an integral part of everyone's lives. It is not just a marketing tool or a way of broadcasting news. It not only allows you to hear what people say about you but gives the space for you to share your own opinions on the matters happening and helps us in influencing people in both positive and negative ways. This can have a very significant impact on people and the decisions they make.

One of the consequences of the rapid increase in the number of social media users is the increase in the increase in the inappropriate use of social media by the users.Workshops and collaborative tasks held recently have stimulated projects regarding the identification of hate speech, toxicity, misogyny, sexism, racism, and abusive content (Zampieri et al., 2020). The convenience of accessing information and being a great source of great conversations, it also makes cyber bullying and hate speech possible. Since it allows us to share our view points on everything, Hate speech on transsexual and homosexual people are very common. Transphobia is when people have a deep-rooted negative prejudice about being transgender or non-binary. Homophobia is the aversion or hatred towards people who are homosexual or gay. This has a negative consequence on people belonging to these minority gender groups.

Despite a greater acceptance of sexual variations and same-sex marriage in many places, Homophobia and transphobia still exist widely and is sustained by many religious, political and individual practises. Many studies have presented that around 8 to 9 out of 10 people are subjected to hate speech online the percentage of transgender and homosexual people in it is significantly high (Schmidt and Wiegand, 2017).

Homophobia and transphobia don't end with conversing. It can take a physical form of violence too. Violence is becoming too common on social media platforms and it influences people negatively. Violence in the form of murder, beating or even sexual violence such as molestation is becoming too common (Flores et al., 2022). Social media has a great part in this. It is a powerful tool that can easily influence many people. Many hateful comments such as "Gay people should be killed mercilessly", "Transgender people should be stoned" are becoming too common and greatly influence the current generation. Numerous workshops and collaborative efforts are currently focused on identifying abusive content as well (Chakravarthi, 2023).

Recognising these homophobic and transphobic comments on social media automatically can make it very easy for us to block these immediately

Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 227–232
March 21, 2024 ©2024 Association for Computational Linguistics

(Pamungkas et al., 2023). This tool can flag all the homophobic and transphobic comments and can make the environment inclusive. It can reduce harm and harassment directed at individuals solely based on their sexual orientation or gender identity. Numerous research efforts are underway to identify abusive content in various local languages as well (Chakravarthi et al., 2023). It helps in influencing the social media users positively and helps in reducing homophobia and transphobia around the world.

2 Related Works

Sentiment analysis is a field in which constant works and researches are being carried on. They have many applications on social and e-commerce platforms.

Sharma et al. (2022) has addressed this classification problem by applying the well established deep learning models, including Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) with GloVe embedding, and transformer-based models like Multilingual BERT (Devlin et al., 2018) and IndicBERT (Kakwani et al., 2020). Results Obtained show that the IndicBERT outperforms all the models and was finally used.

Nozza (2022) solved the challenge of data imbalance by introducing a solution involving data augmentation and ensemble modeling. They fine tuned various large language models, including BERT, ROBERTa (Liu et al., 2019) and HATE-BERT (Caselli et al., 2021). A weighted majority vote Is applied to aggregate their predictions.

Abdul Kareem (2023) has employed Transformer based models widely, as it provides better results than the traditional machine learning models. Implementation includes RoBERTa (Liu et al., 2019), DistilBERT (Sanh et al., 2019), and mBERT (Devlin et al., 2018), with a comparative analysis against previous studies. The Results showed that DistilBERT provided better results than RoBERTa and mBERT.

A hypothesis that states that the performance of the models on a newly constructed dataset with limited data will be improved by data augmentation via Pseudo labeling through transliterating the code mixed text to the parent language. Performance of several models were run and tested by Chakravarthi et al. (2022).

The task done by Bhandari and Goyal (2022) involved multi class classification to identify homophobia or transphobia in YouTube comments. The pipeline comprised a transformer-based classification head and data augmentation for oversampling the English dataset, detailed subsequently.

Multiclass classification system done by Wong et al. (2023), utilizes a BERT based model identifying homophobia and transphobia across English, Spanish, Hindi, Malayalam and Tamil. Retraining XLM RoBERTa (Conneau et al., 2019) with relevant social media data, including script-mixed samples, improved performance, especially in Malayalam. Transformer based models are sensitive to register and language-specific retraining, enhancing classification across various conditions.

3 Task and Data Description

The shared task on Homophobia/Transphobia Detection in social media comments at LT-EDI@EACL 2024[1] was created with the task of detecting Homophobia, Transphobia and non-LGBTQ+ content on YouTube comments. The task concentrates more on regional languages so homophobic and transphobic comments were given in Tamil, Telugu, Malayalam, English, Kannada, Gujarati, Hindi, Marathi languages (hom, 2024).

3.1 Dataset

The dataset provided to us were the data derived from YouTube comments. It mainly had 3 categories: “Homophobic”, “Transphobic”, “non-anti-LGBTQ+”. But most of the dataset had a significantly high amount of “non-anti-LGBTQ+” comments. As the data was too imbalanced, the model might have a bias towards the third category as it had a significantly higher amount of data. So the data had to be up sampled so that the amount of data in each category is equal.

3.2 Up-sampling data

The resample function of "sklearn.utils"[2] is utilized for up-sampling data. It resamples arrays or sparse matrices in a consistent way and the default strategy implements one step of the bootstrapping procedure. It takes feature matrix and corresponding target labels as input. it primarily focuses on minority class subset for up-sampling in order to eliminate any bias. the function randomly selects samples with replacement, potentially duplicating

[1] Fourth workshop on Language Technology for Equality, Diversity, Inclusion (LT-EDI-2024) AT EACL 2024

[2] https://scikit-learn.org/stable/modules/classes.htmlmodule-sklearn.utils

	Training data before Up-Sampling			Training data after Up-Sampling		
Languages	Non anti LGBTQ+	Homophobic	Transphobic	Non anti LGBTQ+	Homophobic	Transphobic
Tamil	2,064	453	145	2,064	2,064	2,064
Hindi	2,423	92	45	2,423	2,423	2,423
Gujarati	3,848	2,267	2,004	3,848	3,848	3,848
Kannada	4,463	2,835	2,765	4,463	4,463	4,463
Marathi	2,572	551	377	2,572	2,572	2,572
English	2,978	179	7	2,978	2,978	2,978
Malayalam	2,468	476	170	2,468	2,468	2,468
Telugu	3,496	2,907	2,647	3,496	3,496	3,496

	Development data before Up-Sampling			Development data after Up-Sampling		
Languages	Non anti LGBTQ+	Homophobic	Transphobic	Non anti LGBTQ+	Homophobic	Transphobic
Tamil	507	118	41	507	507	507
Hindi	305	13	2	305	305	305
Gujarati	788	498	454	788	788	788
Kannada	955	617	585	955	955	955
Marathi	541	129	80	541	541	541
English	748	42	2	748	748	748
Malayalam	937	197	79	937	937	937
Telugu	747	605	588	747	747	747

Table 1: Data before and after Up-Sampling

samples in some cases. It returns the up-sampled feature matrix and target labels, augmenting the original dataset to enhance minority class representation. Table 1 shows training and development before and after up sampling.

4 Methodology

We used many traditional models to test our model from it. We ran all the dataset through Logistic Regression, Support Vector Machine, Random Forest Classifier, K-nearest Neighbour. By noticing the accuracy and the F1 score of each output, we determined if the model was over or under fitting and by comparing all the metrics, we selected the best model out of all the available options. Support Vector Machine (SVM) produced the best output in majority cases.

4.1 Data Preprocessing

The data entered is not of very high quality as it has many unwanted elements in it. So the data undergoes several processes before it is fed into the model. This removes all the insignificant things from our data and makes it ready to be fed into the model

1. The entered data has many words in the upper- and lower-case words. Lower casing in text preprocessing ensures uniformity and simplifies the analysis for the model. This enhances the model's performance.

2. The text entered is filled with a lot of punctuation and emojis. These elements don't add meaning to the sentence. Removing emojis and punctuation in a dataset simplifies the analysis, reduces the noise and also ensures a consistent processing by the model.

3. Stop words are the commonly used words in a language. These are the words that are present highly in any dataset but carry very little useful information for a classification model. As the frequency of these words are too high, it is important to remove these words from the dataset and this results in a smaller data. The stop words for all the languages were downloaded from public repositories and from the "nltk" documentation[3].

[3] https://www.nltk.org/

4.2 Feature Extraction

Most of the Machine learning and Deep learning algorithms are not capable of processing strings or plain text in their raw form. So, we need to feed in numerical numbers as inputs to perform any task. In simple terms, word embeddings are the texts converted into numbers and there may be different numerical representations of the same text.

We employed the TF-IDF vectorizer for this. TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. It is a measure of originality of a word by comparing the number of times a word appears in document with the number of documents the word appears in.

$$TF = \frac{frequency\, of\, term\, in\, the\, document}{total\, number\, of\, terms\, in\, the\, document}$$

$$IDF = \frac{number\, of\, documents\, in\, the\, corpus}{number\, of\, documents\, with\, the\, term}$$

$$TF - IDF = TF * IDF$$

4.3 Classification using ML Models

To classify the data into different categories, we implemented many traditional models for this. the models include SVM, Random Forest, K-nearest neighbour, Logistic Regression as well as some simple transformer models like LaBSE (Feng et al., 2020) and some language specific models like Hindi BERT (Joshi, 2022), Tamil BERT, Telugu BERT, Malayalam BERT, Gujarati BERT, Kannada BERT, bert-base-uncased (Devlin et al., 2018). We noticed that in almost all the datasets, traditional models gave a very high accuracy. In all the cases, SVM gave the highest accuracy and macro average F1 score. Support Vector Machine is one of the most popular Supervised Learning algorithms which is used for classification as well as regression problems. SVM works by mapping data to a high dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable.

4.4 SVM Parameters

The regularization parameter, "C", is set to 1. this parameter helps the SVM optimization in determining the balance between margin size and misclassification. Higher C prioritizes accurate classification, favoring smaller margin hyper planes and lower C seeks larger margin hyper planes, even if it leads to more misclassifications. The kernel configuration is configured as "linear". A linear hyperplane determines the dot product between input vectors in the initial feature space. This dot product calculates the similarity or distance within the original feature space, resulting in a linear hyperplane decision boundary that distinguishes between classes. Verbose, when set to true, serves as an option to display detailed progress updates including epoch, percentage completion, batch processing and estimated time remaining. The probability parameter allows the model to provide class probability estimates rather than just class predictions, enhancing the model's results. All remaining parameters are set to their default values.

5 Results

In every case, the Support Vector Machine model has demonstrated a better performance compared to all other models. The evaluation was conducted by assessing the classification report of training,development and test datasets. The outcomes of Support Vector Machine (SVM) are presented in Table 3, while the results for Logistic Regression

	Training data		**Development data**	
Languages	**Accuracy**	**Macro Average F1**	**Accuracy**	**Macro Average F1**
Tamil	0.88	0.69	0.85	0.56
Hindi	0.95	0.32	0.95	0.33
Gujarati	0.93	0.93	0.93	0.93
Kannada	0.91	0.91	0.91	0.91
Marathi	0.79	0.52	0.78	0.49
English	0.94	0.36	0.94	0.32
Malayalam	0.91	0.76	0.87	0.62
Telugu	0.92	0.92	0.93	0.93

Table 2: Classification report and results for all languages using Logistic Regression

	Training data		Development data		Testing data	
Languages	**Accuracy**	**Macro Average F1**	**Accuracy**	**Macro Average F1**	**Macro Average F1**	**Rank**
Tamil	0.71	0.81	0.89	0.77	0.48	6
Hindi	0.95	0.38	0.97	0.48	0.32	4
Gujarati	0.94	0.94	0.94	0.94	0.89	6
Kannada	0.92	0.92	0.93	0.93	0.88	8
Marathi	0.86	0.70	0.88	0.78	0.39	6
English	0.96	0.57	0.96	0.44	0.34	9
Malayalam	0.97	0.94	0.94	0.88	0.87	5
Telugu	0.93	0.93	0.94	0.94	0.89	6

Table 3: Classification report and results for all languages using SVM (Most Optimum)

are displayed in Table 2.

6 Conclusion

In conclusion, our study delved into the comprehensive evaluation of traditional machine learning models for text classification, employing an array of techniques from Logistic regression to sophisticated models like Support Vector Machine (SVM) and transformer-based approaches. Rigorous preprocessing, including lower casing, removal of punctuation, emojis, and stop words, ensured data quality. The TF-IDF vectorizer facilitated effective feature extraction, translating textual data into numerical representations. Notably, SVM consistently outperformed other models in terms of accuracy and F1 score across diverse datasets. While our traditional models exhibited commendable performance, it is imperative to acknowledge the evolving landscape of deep learning and advanced embeddings, suggesting avenues for future exploration and refinement of models to capture intricate language nuances and patterns.

7 Limitations

Despite the robust performance of traditional machine learning models, our methodology has inherent limitations. The preprocessing steps, while essential for enhancing data quality, may inadvertently lead to information loss. Removing punctuations, emojis, and stop words, although beneficial for noise reduction, could result in the omission of nuanced context. Additionally, the reliance on TF-IDF for feature extraction may not capture complex semantic relationships in the data. While Support Vector Machine (SVM) emerged as a superior model, its effectiveness might be constrained by non-linearly separable data. Furthermore, our approach predominantly focuses on traditional models, potentially overlooking the nuanced representations that more advanced neural networks and embeddings could offer, limiting the model's adaptability to intricate language patterns and contexts.

8 Ethical Statement

While creating the paper, we made sure that the ACL Code of Ethics was practiced throughout the process of working on the Shared Task. This research task was done with the idea of making social media platform a safe space for people regardless of their sexual orientation. It was made sure that credit has been given to all authors whose works and ideas have been used or incorporated in the reference section. The solution proposed follows all the local, regional and international laws and regulations. This solution gives a lot of importance on data privacy, we ensured that no access to data is granted to unauthorized individuals or organisations, thus preventing any leakage of information.

References

2024. Overview of Third Shared Task on Homophobia and Transphobia Detection in Social Media Comments, author = Chakravarthi, Bharathi Raja and Kumaresan, Prasanna Kumar and Priyadharshini, Ruba and Buitelaar, Paul and Hegde, Asha and Shashirekha, Hosahalli Lakshmaiah and Rajiakodi, Saranya and García-Cumbreras, Miguel Ángel and Jiménez-Zafra, Salud María and García-Díaz, José Antonio and Valencia-García, Rafael and Ponnusamy, Kishore Kumar and Shetty, Poorvi and García-Baena, Daniel. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity and Inclusion*, Malta. European Chapter of the Association for Computational Linguistics.

Syed Ebrahim Abdul Kareem. 2023. *Leveraging Transfer Learning Techniques for Homophobia and Transphobia Detection*. Ph.D. thesis, Dublin, National College of Ireland.

Vitthal Bhandari and Poonam Goyal. 2022. bitsa_nlp@ LT-EDI-ACL2022: Leveraging pretrained language models for detecting homophobia and transphobia in social media comments. *arXiv preprint arXiv:2203.14267*.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

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, 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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised Cross-lingual Representation Learning at Scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT Sentence Embedding. *CoRR*, abs/2007.01852.

Andrew R Flores, Rebecca L Stotzer, Ilan H Meyer, and Lynn L Langton. 2022. Hate crimes against lgbt people: National crime victimization survey, 2017-2019. *PLoS one*, 17(12):e0279363.

Raviraj Joshi. 2022. L3Cube-HindBERT and DevBERT: Pre-Trained BERT Transformer models for Devanagari based Hindi and Marathi Languages. *arXiv preprint arXiv:2211.11418*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Debora Nozza. 2022. Nozza@LT-EDI-ACL2022: Ensemble modeling for homophobia and transphobia detection. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 258–264, Dublin, Ireland. Association for Computational Linguistics.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2023. Towards multidomain and multilingual abusive language detection: a survey. *Personal and Ubiquitous Computing*, 27(1):17–43.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Deepawali Sharma, Vedika Gupta, and Vivek Kumar Singh. 2022. Detection of homophobia & transphobia in Malayalam and Tamil: Exploring deep learning methods. In *International Conference on Advanced Network Technologies and Intelligent Computing*, pages 217–226. Springer.

Sidney Wong, Matthew Durward, Benjamin Adams, and Jonathan Dunn. 2023. cantnlp@ LT-EDI-2023: Homophobia/Transphobia Detection in Social Media Comments using Spatio-Temporally Retrained Language Models. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 103–108.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.