SSNTech2@LT-EDI-RANLP-2023: Homophobia/Transphobia Detection in Social Media Comments Using Linear Classification Techniques

Vaidhegi D, Priya M, Rajalakshmi S, Angel Deborah S, Mirnalinee T T
Department of Computer Science and Engineering,
Sri Sivasubramaniya Nadar College of Engineering, Chennai - 603110, Tamil Nadu, India
vaidhegi2110337@ssn.edu.in, priya2110680@ssn.edu.in,
rajalakshmis@ssn.edu.in, angeldeborahs@ssn.edu.in,
mirnalineett@ssn.edu.in

Abstract

The abusive content on social media networks is causing destructive effects on the mental well-being of online users. Homophobia refers to the fear, negative attitudes and feeling towards homosexuality. Transphobia refers to negative attitudes, hatred and prejudice towards transsexual people. Even though, some parts of the society have started to accept homosexuality and transsexuality, there are still a large set of the population opposing it. Hate speech targeting LGBTQ+ individuals, known as homophobia/transphobia speech, has become a growing concern. This has led to a toxic and unwelcoming environment for LGBTQ+ people on online platforms. This poses a significant societal issue, hindering the progress of equality, diversity, and inclusion. The identification of homophobic and transphobic comments on social media platforms plays a crucial role in creating a safer environment for all social media users. In order to accomplish this, we built a ML model using SGD and SVM classifier. Our approach yielded promising results, with a weighted F1-score of 0.95 on the English dataset and we secured 4^{th} rank in this task.

1 Introduction

In the recent years, people have started to discover themselves and open up about homosexuality and trans-sexuality. People have the complete freedom to choose their sexuality from the different choices. The widespread use of social media platforms has revolutionized communication and provided a space for individuals to express their thoughts, opinions, and experiences. The development of social media and support by the respective communities have helped in providing the people the courage to express themselves. However, this has also led to the increase in hate speech and discriminating comments towards this set of people. Even after several laws and regulations, LGBTQIA+ communities are constantly facing violations against them in various forms.

Homophobia and transphobia, which involve negative attitudes, prejudices, and discriminatory behaviors towards LGBTQIA+ individuals, have gained prominence as pressing issues within online platforms. The proliferation of such toxic comments not only inflicts harm upon the targeted individuals but also fosters an environment of fear, intolerance, and exclusion. In response to this escalating problem, the utilization of Machine Learning (ML) techniques has garnered significant attention. ML algorithms offer the potential to automatically detect and classify homophobic and transphobic content in social media comments, enabling swift intervention and mitigating the detrimental impact on the affected individuals. This research paper aims to provide a comprehensive study of the application of ML algorithms for identifying homophobic and transphobic comments on social media platforms. The primary objective is to develop an efficient and accurate model that can autonomously identify and flag such discriminatory content, contributing to the establishment of a safer and more inclusive online environment for LGBTQIA+ individuals.

This paper involves analysing different approaches for classifying the English dataset of social media comments into three categories: Homophobic, Transphobic, and Non-anti-LGBT+ content, as part of the shared task Homophobia/Transphobia Detection @LT-EDI@RANLP-2023. Furthermore, we will discuss about the various methodologies used to process the data, implement the ML model and finally we will take a look into the outcome of the model and future developments.

LT-EDI 2023 – Third Workshop on Language Technology for Equality, Diversity and Inclusion,
pages 166–171, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-084-7_025

2 Related Work

Research on online hate speech has predominantly centered around aggression, sexism, racism, offensive language, hatred, and harassment, with limited emphasis on identifying specific instances of homophobic and transphobic speech (11). Noteworthy studies include an examination of linguistic patterns among homosexual individuals in China using a created corpus (4). Emotion lexicons have been developed to discern acceptable and unacceptable discourse concerning LGBTQ+ topics in languages such as English, Croatian, Dutch, and Slovene (3).

A study analyzed hate speech comments related to LGBTQ+ issues on Facebook, highlighting prevalent themes like repulsion towards the LGBTQ+ community and discrediting of journalistic information (2). Furthermore, a manually annotated corpus was compiled from YouTube, encompassing homophobic and transphobic speech in multiple languages, including English, Tamil, and code-mixed Tamil-English, with the aim of categorizing speech at various levels (14). (12) used the BERT model to detect offensive language as a first level of identification of abusive content.

While psychological studies have examined aspects like homophobic bullying and the impact on mental health, there remains a requirement for empirical evidence and theoretical comprehension of online homophobia/transphobia and its association with Internet usage (5; 7).

Regarding hate speech detection in code-mixed settings, researchers have started to extract code-mixed data from social media platforms due to increased user engagement. However, countries like India, with multiple languages spoken, face a scarcity of pertinent data in low-resourced languages (8). (10) explored the significance of multi-task learning for identifying offensive language and performing sentiment analysis in closely related code-mixed languages like Kannada, Malayalam, and Tamil. Sivanaiah et al. (6) used a deep learning model to identify misogynous content against women using multimodal data.

Biradar et al. (12) (2022) utilized translational systems to convert texts to English and fine-tuned language models for hate speech classification in Hinglish (Hindi-English). However, these approaches may not fully capture contextual nuances and accurately interpret sarcasm. Additionally, pre-trained language models encounter challenges in capturing contextual relationships between code-mixed languages (9).

In summary, although research exists on identifying and addressing online hate speech, there is a need for more focused investigations into homophobic and transphobic speech. Additionally, effectively detecting hate speech in code-mixed settings requires further attention to develop precise and context-aware classification models.

3 Dataset used

The dataset provided in task A of the homophobia/transphobia comments detection in the LT-EDI@RANLP-2023 is used in this work. The dataset consists of 3164 entries under 3 labels namely Homophobia, Transphobia and Non-anti-LGBT+ content. The number of entries under each dataset are 179 for homophobia, 8 for transphobia and 2978 for Non-anti-LGBT+ content. Table 3 shows the data distribution of training dataset and Table 4 shows the distribution of development dataset for various classes.

S.no	Labels	Counts
1	Non-anti-LGBT+content	2978
2	Homophobia	179
3	Transphobia	8

Table 1: Number of classes in train dataset

S.no	Labels	Counts
1	Non-anti-LGBT+content	748
2	Homophobia	42
3	Transphobia	2

Table 2: Number of classes in development dataset

4 Methodologies

We have experimented with multi-class linear classifier ML algorithms for classifying the text into various category of the output class labels. The algorithms are explained in the following sections.

4.1 Stochastic Gradient Descent Classifier

The Stochastic Gradient Descent (SGD) classifier is a highly popular ML algorithm specifically designed for classification tasks. As a member of the linear classifiers family, it is widely recognized for its efficiency and effectiveness, particularly in large-scale learning scenarios. The SGD classifier lever-

ages the optimization technique of stochastic gradient descent, a widely-used iterative approach that minimizes a loss function to find the optimal model parameters. While its primary application is binary classification, it can be extended to handle multiclass problems using techniques like one-vs-all or softmax regression. The algorithm shines when working with sparse data and high-dimensional feature spaces, demonstrating its capability to handle complex datasets.

Versatility and flexibility are among the key advantages of the SGD classifier. It finds application across a diverse range of classification tasks, including text categorization, sentiment analysis, and image classification. Furthermore, its ability to support online learning enables incremental updates to the model as new data arrives, making it suitable for real-time or streaming data scenarios. In summary, the SGD classifier stands out as a flexible and efficient algorithm for both binary and multiclass classification. With its stochastic optimization technique and adaptability to large datasets, it proves valuable in a wide array of applications.

4.2 Support Vector Machine classifier

The Support Vector Machine (SVM) classifier is a widely used and powerful supervised ML algorithm that excels in both classification and regression tasks. Renowned for its effectiveness in managing complex datasets and achieving high accuracy, SVM identifies an optimal hyperplane that separates distinct classes within the feature space. This hyperplane selection maximizes the margin, which denotes the distance between the hyperplane and the nearest data points from each class (known as support vectors). Consequently, SVM exhibits excellent generalization performance, even when confronted with non-linearly separable data.

An advantageous feature of SVM is its efficient handling of high-dimensional feature spaces, rendering it suitable for datasets encompassing numerous features, such as text classification and image recognition. SVM's prevalence stems from its exceptional generalization capabilities and resilience to outliers. By employing appropriate kernel functions, it can handle both linearly separable and non-linearly separable data. As a result, SVMs find application across diverse domains, including text classification, image classification, bioinformatics, and finance. In summary, the SVM classifier stands as a potent ML algorithm that determines an optimal hyperplane to distinguish between classes within the feature space. By leveraging kernel functions to transform data into higher-dimensional spaces, SVM enables effective classification. Its versatility and wide-ranging applications arise from its ability to handle intricate datasets while achieving remarkable accuracy.

5 Implementation

The above model implementation involves the following methodologies:

The following diagram provides a visual flowchart of the implementation steps.

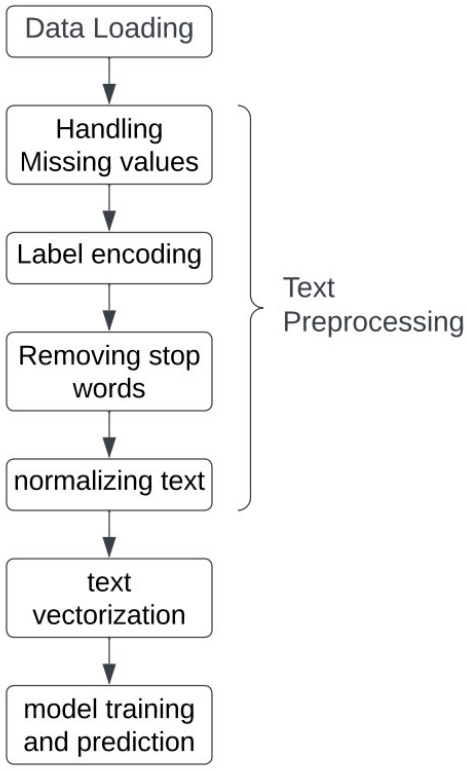

Figure 1: Flow chart of implementation

5.1 Data loading

First step is to import the neccessary libraries. The required libraries such as Pandas, Numpy, Matplotlib, Re, and Sklearn are imported. The dataset is loaded using the read_csv() function from the Pandas library and is converted to PANDAS dataframe. The column names of the DataFrame are renamed using the rename() function.

5.2 Text preprocessing

Various text preprocessing steps are performed using the NLTK library and regular expressions.

5.2.1 Handling missing values

The presence of missing values in the DataFrame is checked using the isna().any() function.

5.2.2 Label encoding

The categorical target variable is encoded using the LabelEncoder from sklearn.preprocessing and each label is given a numerical value.

5.2.3 Removing stop words

Stopwords are words that do not add to the overall meaning of the text. The general stopwords are removed by using functions in NLTK. The NLTK corpus is used to download the required stopwords. The remove_stop_words() function is defined to remove stop words from the text column. The function tokenizes the text, filters out stop words, and joins the filtered tokens back into a string. The stop words are removed from the text column using the apply() function.

5.2.4 Removing custom stop words

Additionally, some common stopwords like *you*, *your* were in short forms like *u*, *ur* as the dataset contains social media comments which can be informal. Therefore, they were not removed by standard nltk library. So, the custom stop words used as short forms which are frequently found in the given dataset are defined in the custom_stop_words set. The remove_custom_stop_words() function is defined to remove these custom stop words from the text column. Similar to the previous step, the function tokenizes the text, filters out custom stop words, and joins the filtered tokens back into a string. The custom stop words are removed from the text column using the apply() function.

5.2.5 Normalizing text

The NLTK tokenizer and WordNetLemmatizer are used to normalize the text. The normalize_text_nltk() function is defined to tokenize the text, lemmatize the tokens, convert them to lowercase, and join them back into a string. The text normalization is applied to the text column using the apply() function.

5.2.6 Removing numerical values

As the numerical data does not contribute to categorizing the text, they are removed. Regular expressions are used to remove numerical values from the text column. The lambda function is applied to each text value using the apply() function and RE.sub() is used to substitute numerical values with an empty string.

5.3 Text vectorization

The TfidfVectorizer in sklearn.feature_extraction .text is used to convert the text data into numerical vectors. The fit_transform() function is used on the training set, and transform() is used on the development and testing set.

5.4 Model training and prediction

Two test runs were done using two ML models, one with SGD classifier and another with SVM classifier.

5.4.1 The SGD classifier

The SGDClassifier from sklearn.linear_model is instantiated and trained using the training data. The trained model is used to predict the target variable for the testing data.

5.4.2 SVM classifier

The Linear SVC (SVM) classifier from sklearn.svm is instantiated and trained using the training data. The trained model is used to predict the target variable for the testing data.

5.5 Evaluation

The classification_report from sklearn.metrics is used to evaluate the performance of the model by comparing the predicted target values with the actual target values of development dataset. The various performance scores like accuracy, macro and weighted precision, macro and weighted recall and macro and weighted f-1 scores of both the models are tabulated in Table 5 and Table 6. Then, the model was run for test dataset and the predicted results were submitted. The evaluation of test dataset was based on weighted f-1 score.

S.no	Labels	Counts
1	Non-anti-LGBT+content	2978
2	Homophobia	179
3	Transphobia	8

Table 3: Number of classes in train dataset

6 Results and Discussion

This task is evaluated on the weighted averages and macro averages of three performance metrics - Precision, Recall and F1-score. The scores for these metrics and the accuracy score achieved for

S.no	Labels	Counts
1	Non-anti-LGBT+content	748
2	Homophobia	42
3	Transphobia	2

Table 4: Number of classes in development dataset

the training and development dataset of Homophobia/Transphobia Detection task under SGD classifier are tabulated in Table 5. The scores achieved for training and development dataset of this task under SVM classifier are tabulated in Table 6.

Metrics	Train DS	Dev DS
Accuracy	0.95	0.94
Macro Precision	0.47	0.57
Macro Recall	0.40	0.55
Macro F1-score	0.43	0.57
Weighted precision	0.94	0.94
Weighted recall	0.95	0.94
Weighted F1-score	0.94	0.91

Table 5: Performance of SGD classifier for training and development dataset

Metrics	Train DS	Dev DS
Accuracy	0.96	0.93
Macro Precision	0.56	0.47
Macro Recall	0.39	0.50
Macro F1-score	0.42	0.48
Weighted precision	0.95	0.87
Weighted recall	0.96	0.93
Weighted F1-score	0.95	0.90

Table 6: Performance of SVM classifier for training and development dataset

It is inferred that the transphobia and homophobia samples are much smaller than the non-anti-lgbt+ content samples. Therefore the individual accuracy and F1 scores for the categories also vary. It is inferred from the experimental results that SVM is not performing well when compared to SGD classifier as the data is highly imbalanced across the various classes. Our submission using SGD model secured the 4^{th} rank in Task A, i.e., Homophobia / Transphobia Detection on English dataset. Our model procured a weighted F1-score of 0.9582 while the top rank team secured 0.969 score. Thus, SGD has been an effective model for the test data also when compared to SVM.

As the given dataset was skewed, we tried for data augmentation. But this method was not effective. Since we had less number of samples in the dataset we used machine learning techniques such as SVM classifier and SGD classifier, as deep learning techniques require large amount of data to learn.

7 Conclusion And Future Works

In this research, we conducted a comparative analysis of various models for the shared task on homophobia/transphobia detection at LT-EDI@RANLP-2023. Our findings revealed that the SGD Classifier yielded the most favourable results for English text. The current dataset was skewed and the prediction scores were low for transphobia text. We may further train the model with enhanced datasets with more data and category labels to get more accuracy. The model was trained for monolingual text (English). Extending this, a machine model to detect multilingual text can be built. We are also aiming to investigate on the data augmentation techniques for this specific case.

References

[1] Ashraf, N., Taha, M., Abd Elfattah, A., & Nayel, H. (2022, May). Nayel@ lt-edi-acl2022: Homophobia/transphobia detection for Equality, Diversity, and Inclusion using SVM. In Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion (pp. 287-290).

[2] Silva, M. P. D., & Silva, L. S. D. (2021). Hate speech dissemination in news comments: analysis of news about LGBT universe on Facebook cybermedia from Mato Grosso do Sul. Intercom: Revista Brasileira de Ciências da Comunicação, 44, 137-155.

[3] Ljubešić, N., Markov, I., Fišer, D., & Daelemans, W. (2020). The lilah emotion lexicon of Croatian, Dutch, and Slovene. In Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (pp. 153–157). Barcelona, Spain (Online), ACL.

[4] Wu, H. H., & Hsieh, S. K. (2017, November). Exploring Lavender Tongue from Social Media Texts [In Chinese]. In Proceedings of the 29th Conference on Computational Linguistics and Speech Processing (ROCLING 2017) (pp. 68-80).

[5] Bacchini, D., Esposito, C., Affuso, G., & Amodeo, A. L. (2021). The impact of personal values, gender stereotypes, and school climate on homophobic bullying: a multilevel analysis. Sexuality Research and Social Policy, 18, 598-611.

[6] Ventriglio, A., Castaldelli-Maia, J. M., Torales, J., De Berardis, D., & Bhugra, D. (2021). Homophobia and mental health: a scourge of the modern era. Epidemiology and Psychiatric Sciences, 30, e52.

[7] Roy, P. K., Bhawal, S., & Subalalitha, C. N. (2022). Hate speech and offensive language detection in Dravidian languages using deep ensemble framework. Computer Speech & Language, 75, 101386.

[8] Hande, A., Hegde, S. U., & Chakravarthi, B. R. (2022). Multi-task learning in under-resourced Dravidian languages. Journal of Data, Information and Management, 4(2), 137-165.

[9] Yasaswini, K., Puranik, K., Hande, A., Priyadharshini, R., Thavareesan, S., & Chakravarthi, B. R. (2021, April). IIITT@ DravidianLangTech-EACL2021: Transfer learning for offensive language detection in Dravidian languages. In Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages (pp. 187-194).

[10] Biradar, S., Saumya, S., & Chauhan, A. (2022). Fighting hate speech from a bilingual Hinglish speaker's perspective: a transformer-and translation-based approach. Social Network Analysis and Mining, 12(1), 87.

[11] Kumar, G., Singh, J. P., & Kumar, A. (2021). A deep multi-modal neural network for the identification of hate speech from social media. In Responsible AI and Analytics for an Ethical and Inclusive Digitized Society: 20th IFIP WG 6.11 Conference on e-Business, e-Services and e-Society, I3E 2021, Galway, Ireland, September 1–3, 2021, Proceedings 20 (pp. 670-680). Springer International Publishing.

[12] Chakravarthi, B. R., Priyadharshini, R., Durairaj, T., McCrae, J. P., Buitelaar, P., Kumaresan, P., & Ponnusamy, R. (2022, May). Overview of the shared task on homophobia and transphobia detection in social media comments. In Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion (pp. 369-377).

[13] Sivanaiah, R., Angel, S., Rajendram, S. M., & Mirnalinee, T. T. (2022, July). TechSSN at semeval-2022 task 5: Multimedia automatic misogyny identification using deep learning models. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022) (pp. 571-574).

[14] Sivanaiah, R., Suseelan, A., Rajendram, S. M., & Tt, M. (2020, December). TECHSSN at SemEval-2020 Task 12: Offensive language detection using BERT embeddings. In Proceedings of the Fourteenth Workshop on Semantic Evaluation (pp. 2190-2196).

[15] Chakravarthi, B. R., Hande, A., Ponnusamy, R., Kumaresan, P. K., & Priyadharshini, R. (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.

[16] Chakravarthi, B. R., Priyadharshini, R., Durairaj, T., McCrae, J. P., Buitelaar, P., Kumaresan, P., & Ponnusamy, R. (2022, May). Overview of the shared task on homophobia and transphobia detection in social media comments. In Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion (pp. 369-377).