Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230958

Detecting homophobic and transphobic texts from youtube comments using machine learning models

Malliga Subramanian[1, 3], Veerappampalayam Easwaramoorty Sathishkumar[2], Kogilavani Shanmugavadivel[1], P Deva[1], S Haris[1], and Jaehyuk Cho[2]

[1]Department of Computer Science and Engineering Kongu Engineering College, Perundurai, Tamilnadu, India
[2]Department of Software Engineering, Jeonbuk National University, Jeonji-si, Jeollabuk-do, Republic of Korea

[3]choijh@jbnu.ac.kr

Abstract. Today's world has witnessed an exponential rise in disseminating degrading and offensive content via social media. A global increase in violence against minorities, such as gun violence, murders, and forced displacement, has been connected to using harsh and derogatory language online. The policies enacted to prevent abusive or derogatory language risk stifling free speech and are applied differently. These languages can affect the mental state of social media users. Homophobic and transphobic expressions are a insulting people's sexuality or character among harsh and disrespectful remarks. To study social media information and discriminate between homophobic and transphobic comments, it is necessary to construct language-based automatic categorization methods. This dataset consists of Tamil YouTube comments collected as part of the Shared Task on Sentiment Analysis and Homophobia detection of YouTube comments in Code-Mixed Dravidian Languages. This study detects and categorises abusive comments using machine learning models such as SVM, Naive Bayes, Logistic regression, and KNN. In comparison to other classifiers, the accuracy of logistic regression is the highest, at 55%.

Keywords: Homophobia, Transphobic, Social Media, YouTube, Tamil, Machine Learning

1. Introduction

Fortunately, the Internet promotes speedy communication among people around the globe. Technology has made it much simpler to exchange information, seek assistance, and much more, allowing us to remain in touch with friends and family and meet new people worldwide. While most of this relationship is polite and pleasant, others abuse and take advantage of it. Abusive/offensive language is disturbing regardless of whom it is directed towards or who hears it, but it is especially upsetting when it impacts children and adolescents, who may lack the experience or emotional maturity to overcome or absorb it, or the knowledge of where to seek assistance. Some individuals may experience both humiliation and loss of self-respect, which can lead to grief, fury, and antisocial behavior. The abusive and offensive language contains strong terms. It is any communication intended to criticize, insult, or incite hatred against a group or category of people. It may take place in-person, online, or both. It can be sent by text, symbols, pictures, memes, gestures, and video.

DOI: 10.54254/2755-2721/6/20230958

Abusive and offensive speech can be detrimental to people, groups, and society. According to research on the effects of racially, ethnic, gendered, religious, and Lesbian, Gay, Bisexual, Transgender and Queer (LGBTQ) inappropriate language, the targeted may endure substantial emotional, mental and physical damage. These may include low self-esteem, anxiety, life fear, and self-injury. Such language hinders interpersonal connections. It decreases social connection and empathy. Because it typically employs stereotypes and scapegoating, it hinders the ability to address the underlying causes of societal problems.

Homophobia is a broad term that encompasses a variety of negative feelings and attitudes against homosexuals or those who are perceived or recognized as lesbian, gay, or bisexual. It has been described as contempt, prejudice, dislike, hostility, or antipathy, and it may be motivated by irrational fear or have a connection to one's religious views. Transphobia is the term for prejudice against transgender and transsexual persons. An assortment of unfavorable views, sentiments, or behaviors toward transgender people, or transness in general, are collectively referred to as transphobia. Fear, aversion, hostility, aggression, or fury toward anyone who does not fit social gender norms are all examples of transphobia.

The vocabulary used to degrade LGBTQ people includes motifs, buzzwords, and statements that have been employed against homosexuality and other non-heterosexual sexualities. They may be insulting and disrespectful, or they may express hostility toward homosexuality on moral, medical, or even religious grounds. Therefore, rapid and efficient detection and filtering of homophobia and transphobia on the Net will assist in clearing cyber, creating a pleasant , healthy online community, and raising awareness of the unfair treatment of LGBTQ people [1].

In Natural Language Processing (NLP), systems for the automated detection of abusive / offensive language against the LGBTQ communities often fall into one of the following categories: i. classifiers based on features ii) fine-tuned pre-trained language models like as BERT, MuRIL, and RoBERT among others. In this research, machine learning models are used to automatically identify Transphobic, Homophobic, and Non-Anti LGBTQ texts in YouTube user comments. This study focuses primarily on Tamil, a classical Indian language for which data and models are scarce in prior research.

2. Literature survey

With the proliferation of smart gadgets, mobile platforms, and social network environments, these technologies' negative social effects—such as cyberbullying through hurtful comments and rumors—have worsened. Internet trolling has become an unwelcome, global social issue [2]. Once seen as safe havens for information exchange and interpersonal support, online social networks have evolved into hubs for spreading destructive behaviors, political propaganda, and radicalizing content. Toxic users frequently hide behind the guise of anonymity in order to start meaningless discussions and divert other users' attention from a community's primary aims [3]. Along with all the fresh opportunities, there are dangers to be wary of, such as correspondence that contain racial or sexually explicit statements. Automated methods are investigated as solutions since it is impossible to manually monitor and analyze each message. In this section, we provide a brief summary of the research attempts that focused on detecting abusive/offensive comments against the LGBTQ community.

Transphobic/homophobic speaking is a form of inappropriate words that can be summed up as hate speech directed at LGBT+ individuals, and it has become an increasing issue in recent years. Chakravarthi et al. [4] introduced a new hierarchy classification system for internet racism and homophobia, as well as a specialist dataset that allows transphobic/homophobic content to be automatically identified. To encourage the researchers to work on the detection of abusive comments against LGBTQ community, a few shared tasks have been released. Shared tasks are competitions to which researchers submit systems that address specific, predefined challenges. One such shared task has been released as a part of [5]. This task contains the dataset having homophobic and transphobic comments. An overview of the models submitted for the shared task is presented by [6]. Chakravarthi et. al. [6] reviewed the research attempts focused on the shared task and presented an overview. In addition, several research attempts have been carried out to determine the comments against the LGBTQ. Below, we present s brief summary of such attempts.

DOI: 10.54254/2755-2721/6/20230958

The authors of [7] proposed multiple languages Deep Bidirectional Portrayals from Transformers (M-BERT) model that can effectively inform if Turkish comments on social media comprise homophobic or comparable hateful speech. The Homophobic-Abusive Turkish Comments (HATC) set of data was composed of remarks from Instagram that were utilized to prepare the detection systems. Singh et al. [8] created a categorization method that looks at remarks and utilises a Zero-Shot learning framework to find out if they comprise any kind of homosexuality or transphobia. The F1-scores for the Tamil datasets were 0.40, 0.85, and 0.89. Asraf et al. [9] made models using algorithms for machine learning like Svm Algorithm, Random Forest, Passive Aggressive.. Just on data - sets provided as a part of the shared task LT-EDI-ACL2022 [10], they found that SVM was more accurate.

The authors of [11] employed word embeddings, Svm to get an F1-score of 0.93 on the English data, 0.75 on the Tamil set of data, plus 0.87 on the Tamil-English Code-Mixed set of data. [12] employed a RoBERTa-based approach to explore on the task [10] data. Given the substantial minority class, the authors of [13] authors proposed a method using data augmentation and ensembles modelling. They fine-tuned big linguistic models utilised the balanced democratic majority on their predictions. This work scored 0.48 macro and 0.94 weighted F1-score. In machine learning-based systems, a defined preprocessing and feature extraction strategy has always been needed. We employ two feature extraction strategies and transliteration to evaluate model performance.

3. Materials and methods

3.1. Data description

This project's provocative texts adhere to the classic definition of offensive language frequently referred to as "flames," which refers to "offensive messages or remarks that in specific contexts are improper, lack respect for certain groups of people, or are simply unpleasant". Code-Mixed Dravidian YouTube comments are analyzed for homophobia/transphobic and Non-LGBTQ comments. The remarks are written in a combination of Tamil and English (A mixture of Native and Roman Script). In the dataset, there are both Tamil and English comments, not all of which are completely in English but, joined with the Tamil. Even though the average sentence length in the corpus is a single word, the comments in the dataset consist of many sentences. Each of the 2662 training texts and 649 test texts were labeled as Transphobic, Homophobic, or Non-anti-LGBT+. The training set contains 155 transphobic, 2022 non-anti-LGBT+, and 485 homophobic comments. Due to the inconsistency of the dataset, the existing data are supplemented by rearranging the texts to generate new texts. The words of each sentence were altered to form a new phrase. Table 1 contains examples of data set annotations.

Table 1. Sample training data from the dataset (continue).

S. No	Texts	Label
1	ஆமா அண்ணா சரியா சொன்னீங்க govarnmend வேலை குடுத்தா உலகம் தாங்காது	Non-anti-LGBT+ content
2	பண்றது டிச்சிங் டிச்சிங். இதுல நியாய வேற போடி ஏலியன் மண்டை	Homophobic
3	எல்லா ரயிலிலும் இந்த உதவாக்கரைகளின் அட்டகாசம் தாங்கமுடியல.ரயில்வே காவல் துறை நினைத்தால் ஒழித்துவிட முடியாதா.?	Transphobic
4	அவன் யாரு டா அவன் வில்லன் என்ட்ரி குடுக்குறான்,, அவனும் பொட்டயா	Transphobic

DOI: 10.54254/2755-2721/6/20230958

5	Raj Kutty அப்படிய	Non-anti-LGBT+ content

Table 1. (continued).

6	கலாச்சாரத்தின் மீதான பேரிடி....	Homophobic

3.2. Pre-processing and feature extraction

The first stage of building a classifier is to preprocess the dataset (or clean the data). As the corpus contains emojis, punctuation characters, and non-Tamil texts, they were eliminated by preprocessing. The Emoji and Demoji for the Python module to turn the emoticons into text data have also been used. Removal of punctuation is done using the Python string library. To summarize, the following processes have been taken:

- **Removal of emojis:** Emojis and emoticons are used in the text communications in the dataset. They can be fully eliminated or changed to a textual word. Instead, their textual equivalents have been utilized in this study.
- **Removing all punctuation, numerals, and non-Tamil text:** The extra white space, punctuation like!,?, etc., and numerals have been removed in addition to removing non-Tamil texts. These characters make it easier to read, but they are not helpful for determining a person's point of view. Duplicate comments have also been removed.
- **Transliteration :** Then the removal of words which are not in English is performed using the library known as indic_transliteration. A word written in one language is said to be transliterated when it is written using the alphabet of a different language. As the Tamil, Malayalam, and Kannada datasets contain the comments in English, the AI4Bharat Indic-Transliteration engine [14], which covers 21 main Indian languages, is used for transliterating the comments written in English. People unfamiliar with a language's alphabet will find it a little easier to understand when it is transliterated. Table 2 provides a transliteration example.

Table 2. Example for transliteration.

Comments	Language
Suspence lam lilla nanba....pls prank artistku oru request.en friend	Tamil comments transliterated in English
சஸ்பென்ஸ்லாம் லில்லா நண்பா. கலைஞருக்கு ஒரு வேண்டுகோள்.என் நண்பா	Tamil

And, finally, we did the removal of stop words by placing all the stop words in the text file and replacing them from the training dataset. Oversampling is done using random and in-built string methods, and the augmented texts for the category of Transphobic and Homophobic are obtained. For transphobic, the methods are called for 15 times the available data, whereas for homophobic, the methods are called for 3 times the available data.

3.3. Word embeddings

Count Vectorizer and TF-IDF have been used to tokenize each comment. Both extract and represent features from text data. It lets us control n-gram size, custom preprocessing, stop words, tokenization, and vocabulary size. For Count Vectorizer and TF-IDF, the Python scikit-learn library's is a great tool. Count Vectorizer is a very common algorithm to transform the text into a meaningful representation of numbers which is used to fit machine algorithms for prediction. It turns the probability of each phrase in a text into vector. This is useful when working with multiple texts and converting each word into a

DOI: 10.54254/2755-2721/6/20230958

vector and can be used in further text analysis. In TfidfVectorizer, we consider a word's entire document weight. It helps us deal with the most common words. TfidfVectorizer weights the word counts based on their frequency of occurrence in the documents. For the vectorizers, the feature map is obtained. From both feature maps, top-level features are extracted and further given to the various models for training. The output of the vectorizer is an encoded vector, including the vocabulary length and the number of occurrences of each word in the text. Table 3 displays CountVectorizer output samples for the input dataset. These words are not stored as strings by CountVectorizer. They are instead allocated an index value. For instance, "தாங்காது" is assigned index 0, "உலகம்" is assigned index 1, and so on. Using the fit() and transform() functions, the value of each vector for individual text in the training set of data is determined prior to their input to the classifiers.

Table 3. Output from count vectorizer.

Text	தாங்காது	உலகம	கலாச்சாரத்தின்
1	1	1	0
2	0	0	0
3	0	0	1

3.4. Classifiers

After translating the text into word embeddings, the classification models were developed to detect comments against LGBTQ people. On the basis of machine learning, four classifiers are utilized: the Nave Bayes, Logistic Regression, SVM and KNN. These classifiers take as input the preprocessed data and output the categorization of each sentence. Using Scikit-learn Python libraries, the proposed models were implemented.

The Naive Bayes Multinomial Classifier is an improved probabilistic model. Multinomial classifier models count directly, perform well with minimal training data, and learn rapidly. Simple Naive Bayes models the presence or lack of single words within a text. Naive Bayes predicts whether a text contains particular words. Using training data, supervised classification methods build an ideal hyperplane to categorize fresh data. SVM is supervised classification. SVMs can handle plenty of data because they find a hyperplane that optimises class distance. Support vectors provide binary classification by dividing instances into two nonoverlapping categories. SVM classifiers excel at text classification. In our study, we used the RBF kernel function with the scikit-learn SVC() technique. In logistic regression, a function connects one attribute to two or more. This model uses L2 regularisation for binary dependent variables like homophobic, transphobic, or non-anti LGBT+ material. Integrating independent and predictive elements determines a result's probability. KNN classifies new information by comparing it with all already obtained data according to their similarity.

3.5. Model implementation

The classification models have been implemented using the Python programming language and its Sci-kit learn module. The suggested models were trained and tested using Google Colab. The Colab is a browser-based, cloud-based text editor on Jupiter development platform for Python that does not need a desktop setup.

4. Results and discussion

This paper examines the results of classification models developed for automatically identifying hate speech against the LGBTQ+ population in Tamil social media writings. Every classifier is trained with characteristics from the training dataset, then classifiers are then tested with the test dataset supplied.

DOI: 10.54254/2755-2721/6/20230958

4.1. Performance metrics

● Accuracy, Recall, Precision and F1-Score are calculated to find the performance of various models utilized to solve the classification problem. The subsequent section describes these metrics often applied in classifier evaluation.

$$\textbf{\textit{Accuracy}} = (TN + TP) / (TN + TP + FN + FP)$$

Whereas (TP) represents true positives, (TN) represents true negatives, (FP) represents false positives and FN represents false negatives.

● Recall (also known as Responsivity or True Positive Rates) is the percentage of texts properly classified as belonging to a certain class compared to the total number of words that truly belong to that class.

$$\textbf{\textit{Recall}} = (TP+FN) / (TP)$$

● Precision is one metric of the performance of a machine learning model — the quality of a model's accurate prediction, and is given in below equation.

$$\textbf{\textit{Precision}} = TP / (TP+FP)$$

● F1-score is a machine learning measure. It combines accuracy and recall to summarize a model's prediction performance.

$$\textbf{\textit{F1-Score}} = (2*Recall*Precision) / (Recall + Precision)$$

Each classifier is trained with the retrieved features from training set, evaluated with provided test dataset. Table 4 presents the performances of the machine learning models for the word embeddings obtained for Countvectorizer. The confusion matrices for these models are shown in Figure 1. The X axis reflects expected courses, whereas the Y axis shows actual classes. In the table below, class 0 represents homophobic content, class 1 represents non-anti-LGBT+ content, and class 2 represents transphobic content.

Table 4. Performance of classifiers with CountVectorizer.

Classifiers	Class Labels	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Naive Bayes	0	41	27	04	06
	1		44	62	51
	2		39	53	45
KNN	0	39	04	02	02
	1		07	50	12
	2		83	41	55
Logistic Regression	0	55	12	06	08
	1		88	57	69
	2		15	73	25
SVM	0	55	08	06	07
	1		96	57	71
	2		07	89	12

DOI: 10.54254/2755-2721/6/20230958

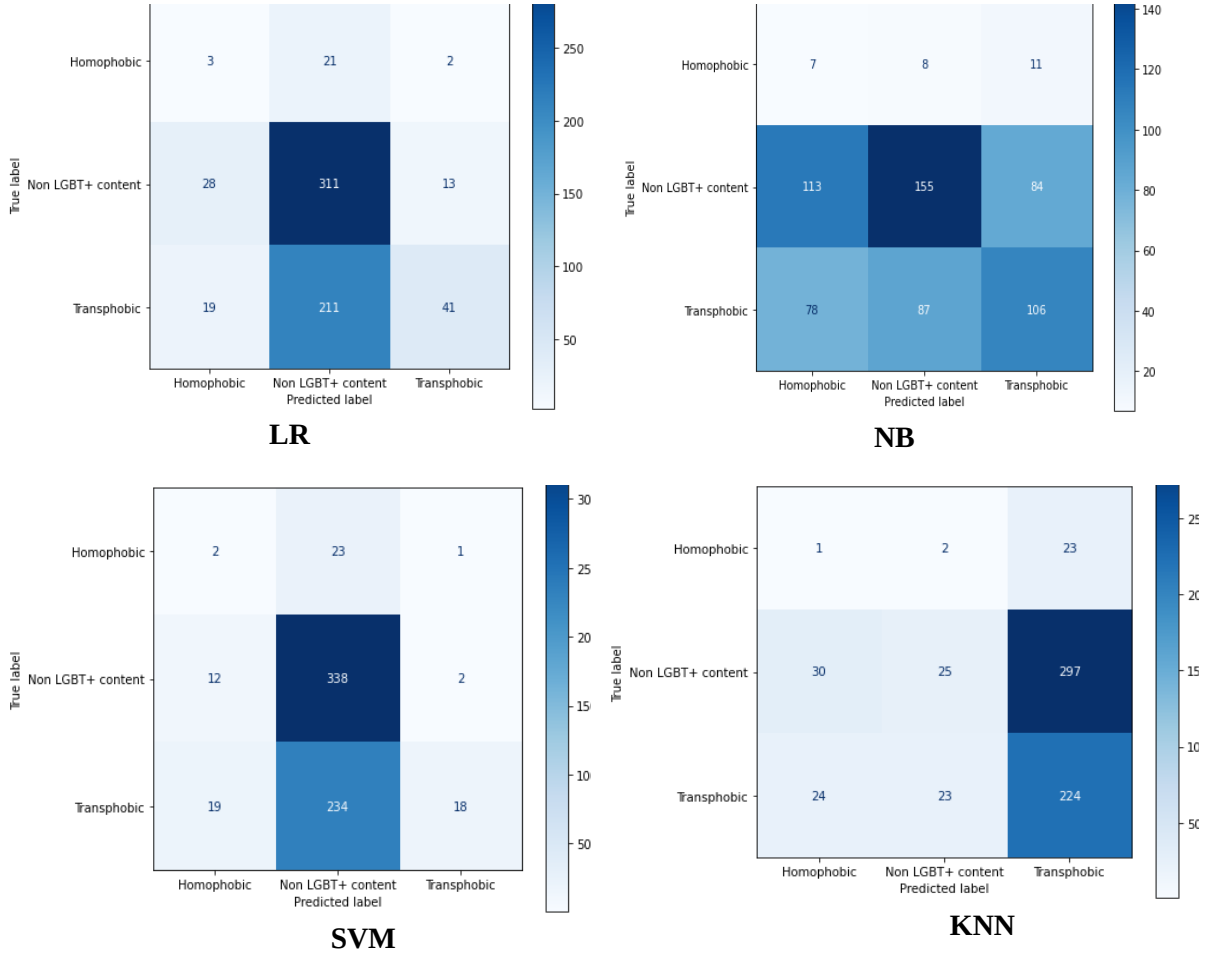

Figure 1. Confusion Matrix of CountVectorizer.

Table 5 presents the performances of the machine learning models for the word embeddings obtained for TF-IDF vectorizer. Figure 2 presents the confusion matrices for these models.

Table 5. Performance of classifiers with TF-IDF.

Classifiers	Class Labels	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Naive Bayes	0	34	31	03	05
	1		34	65	45
	2		33	50	40
KNN	0	31	31	03	05
	1		44	58	50
	2		13	47	21
Logistic Regression	0	39	31	03	06
	1		53	60	56
	2		22	54	31
SVM	0	43	12	02	03
	1		76	56	64
	2		03	50	06

DOI: 10.54254/2755-2721/6/20230958

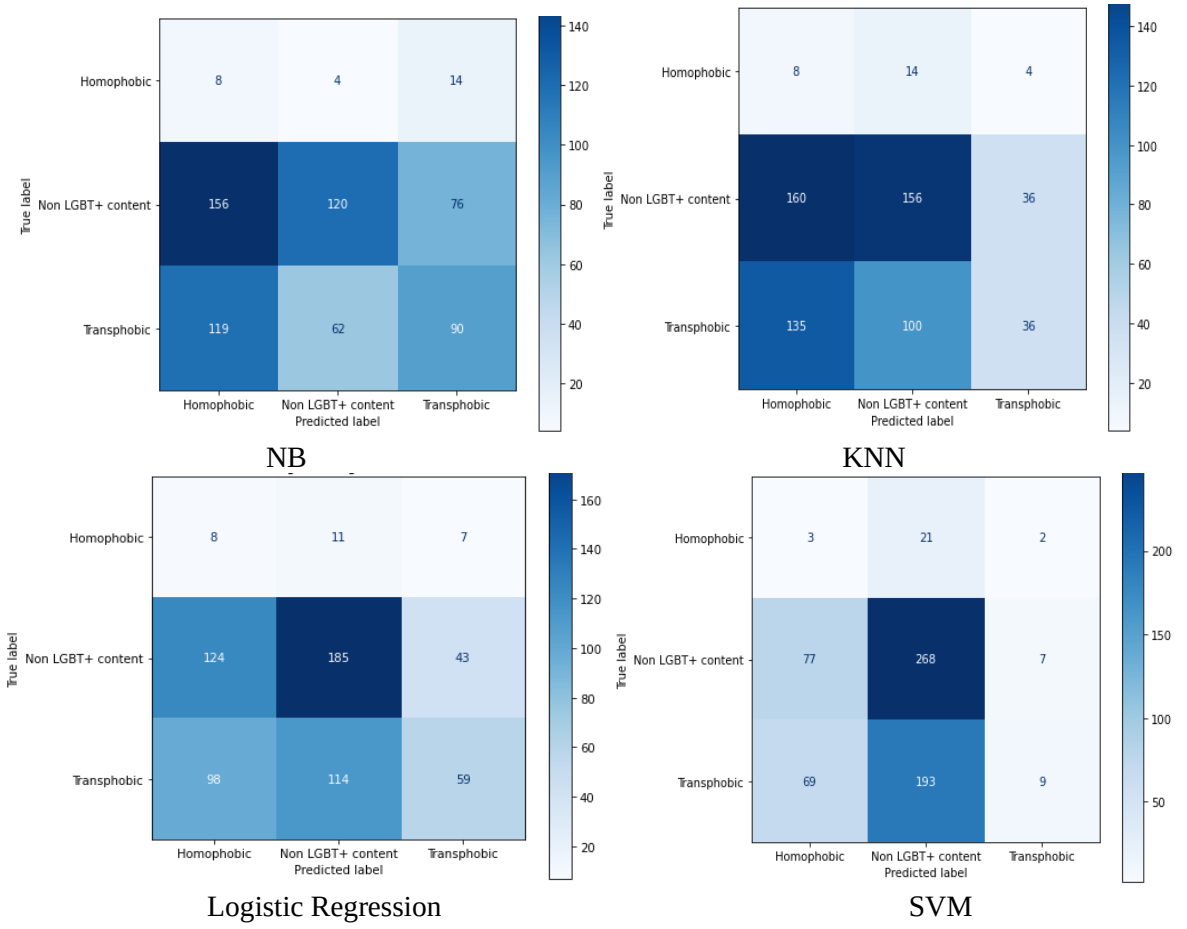

Figure 2. Confusion Matrix for TF-IDF/.

Tables 4 and 5 show that the performance of the classifiers was not appreciable for Transphobic comments compared to the other two classes. Texts for this dataset were supplemented by custom text augmentation methods for oversampling; however, poor scores for class 2 were obtained. Since the number of comments in this class is significantly low compared to the other two classes, we believe this might be one of the reasons for poor performance in this class.

5. Conclusion

This research offered experimental effort and related findings for the challenge of detecting objectionable material in a dataset of code-mixed On order to solve the problem of bad language on social networking, we must study Dravidian language. It provides a comprehensive examination of past offensive text classification methodologies as well as fresh models for automatically recognizing offensive phrases in Tamil. The dataset's features were extracted using Count vectorizer and TF IDF, and numerous classifiers, including SVM, KNN, Naive Bayes and Logistic Regression, were created for this task. Naive Bayes and Logistic Regression both demonstrated more accuracy than the other models. This study might be enhanced by including alternative statistical models of words and new neural network-based algorithms with complex linguistic properties. This initiative adds to studies on identifying objectionable material in languages with scarce funds, like Tamil.

Future directions for work include building the dataset for additional Dravidian languages, which is something interested in doing. Additionally, by crawling and annotating other social media data sets, there is a hope to greatly expand the Tamil data set. Also to increase/enhance the performance of classifiers, this project also intends to investigate semi-supervised approaches as well as incremental

DOI: 10.54254/2755-2721/6/20230958

approaches. Additionally, given manual analysis revealed that many anti-LGBT+ comments are sarcastic, it is intended to investigate the relationship between sarcasm and anti-LGBT+ comments.

Acknowledgments

This research is Supported by the Korea Evaluation Institute of Industrial Technology(KEIT) funded by the Korea Government, Ministry of Trade, Industry and Energy (MOTIE) (Development of Mixed Signal SoC with complex sensor for Smart Home Appliances) under Grant 20010098

References

[1] Moyano N and Mar Sanchez-Fuentes M 2020 Homophobic bullying at schools: A systematic review of research, prevalence, school-related predictors and consequences *Aggression and violent behavior*, vol. 53, p 101441.

[2] Lee S H and Kim H W 2015 Why people post benevolent and malicious comments online *Communications of the ACM*, vol. 58, no. 11, p 74-79.

[3] Obadimu A M 2020 Assessing the Role of Social Media Platforms in the Propagation of Toxicity University of Arkansas at Little Rock.

[4] Chakravarthi B R, Priyadharshini R, Ponnusamy R, Kumaresan P K, Sampath K, Thenmozhi D, Thangasamy S, Nallathambi R, and McCrae J P 2021 Dataset for identification of homophobia and transophobia in multilingual YouTube comments *arXiv preprint arXiv:2109.00227*.

[5] Chakravarthi B R, Muralidaran V, Priyadharshini R, McCrae J P, García M A, Jiménez-Zafra S M, Valencia-García R, Kumaresan R, and Ponnusamy R 2022 Overview of the Shared Task on Hope Speech Detection for Equality, Diversity, and Inclusion *In Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* p 378-388.

[6] Chakravarthi B R, Priyadharshini R, Durairaj T, McCrae J P, Buitelaar P, Kumaresan P, and Ponnusamy R 2022 Overview of The Shared Task on Homophobia and Transphobia Detection in Social Media Comments *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* p 369-377.

[7] Karayiğit H, Akdagli A and Aci C I 2022 Homophobic and hate speech detection using multilingual-BERT model on turkish social media *Information Technology and Control*, vol. 51, no. 2, p 356-375.

[8] Singh M and Motlicek P 2022 IDIAP Submission@ LT-EDI-ACL2022: Homophobia/Transphobia Detection in social media comments *In Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* p 356-361.

[9] Ashraf N, Taha M, Abd Elfattah A, and Nayel H 2022 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* p 287-290.

[10] Chakravarthi B R, Bharathi B, McCrae J P, Zarrouk M, Bali K, and Buitelaar P 2022 Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion. *In Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*

[11] Swaminathan K, Bharathi B, Gayathri G, and Sampath H, 2022 SSNCSE_NLP@ lt-edi-acl2022: Homophobia/transphobia detection in multiple languages using SVM classifiers and Bert-based Transformers *In Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* p. 239-244.

[12] Maimaitituoheti A 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*, p 155-160.

[13] Nozza D 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* p 258-264.

DOI: 10.54254/2755-2721/6/20230958

[14] "AI4Bharat," https://pypi.org/project/ai4bharat-transliteration/