Sentiment and Homophobia Detection on YouTube using Ensemble Machine Learning Techniques

Sunil Saumya, Vanshita Jha and Shankar Biradar

Indian Institute of Information Technology Dharwad
Central University of Rajasthan, India
Indian Institute of Information Technology Dharwad,

Abstract

Internet users frequently express themselves through posts, comments, and articles. The examination of such posts/comments has recently attracted the research community's attention. Sentiment analysis and the identification of homophobic comments are two key research areas in this field. Sentiment analysis reveals that people's emotions reflect positive, negative, or mixed feelings about a certain topic or article. Further, Homophobia refers to a wide range of attitudes and feelings toward people who identify as homosexual, transgender, lesbian, gay, or queer. To encourage research in this direction, the organisers of the Dravidian LangTech shared task as part of FIRE 2022 have set two shared tasks. Task A consists of a message-level polarity detection problem, in which the given YouTube comments system has to recognise positive, negative, and mixed emotions. Task B involves detecting transphobic and homophobic YouTube comments. Our team participated in both subtasks; we worked on the Kannada dataset for sentiment analysis, and our best-performing model secured 11th place among the participating teams. For Task B, we participated in all four languages (Tamil, English, Malayalam, and Tanglish) and received 6, 6, 2, and 4th positions, respectively. In our proposed approach, we employed several Machine learning models, the Ensemble method and Deep learning models to achieve the desired result.

Keywords

Homophobia, Trans phobia, CodeMixed, Ensemble

1. Introduction

Social media websites, blogs, and microblogging sites have become very prominent in today's world, where people can easily share their thoughts and opinions on various real-time scenarios. These websites have also become a source of all kinds of information. Naturally, these comments, posts, and articles tend to infer different things for different people across the world. The comments which are good for some people may not be in the best interest of others. Hence there are various emotions on the same topic, post or issue. These sentiments can be classified into Positive, Negative, Mixed feelings or Unknown states. Analysing each comment, post or article in these categories is known as Sentiment Analysis. Nowadays, sentiment analysis [1] has become very important in various fields like the market, film industry, gaming industry, e-commerce [2] etc. Further, it helps the companies to find the sentiment of people about a particular product or customer needs and understand feedback provided by the customers.

Forum for Information Retrieval Evaluation, December 09-13, 2022, India
sunil.saumya@iiitdwd.ac.in (S. Saumya); vanshitajha@gmail.com (V. Jha); shankar@iiitdwd.ac.in (S. Biradar)

The application of sentiment analysis is present in almost all industries, which can be used to understand the consumers' sentiments and work accordingly.

LGBTQ+ community refers [3] to the group/community of people who identify as lesbian, gay, bisexual, transgender, or queer, all of the gender identities and sexual orientations that are not specifically covered by the other five initials. Homophobia refers to the negative attitude toward people identifying as homosexual, transgender and queer. As a result of homophobia and transphobia, LGBTQ people may face considerable psychological stress, which will prevent them from participating in normal social activities and may potentially result in major mental illness. As a result, quick and effective detection and screening of homophobia and transphobia on the Internet will help to clear cyberspace, create a pleasant and healthy online community, and raise awareness of the unfair treatment of LGBTQ groups [4].

Several studies on sentiment analysis have been undertaken in recent years; however, most of these studies have focused on high-resource languages such as English [5, 6]. Furthermore, relatively very few people have worked on regional south Indian languages [7, 8]. To encourage research on this topic, DravidianLangTech organisers published data in south Indian languages such as Kannada, Tamil, and Malayalam as part of the Fire 2022 proceedings [9]. The shared task organisers provided two subtasks: Task A focuses on sentiment analysis in Kannada language YouTube comments, and Task B focuses on Homophobic comment recognition from social media comments. Our team participated in both challenges and received a good ranking. This article will provide the working notes for our proposed model.

The rest of the article is organized as follows. The next section, 2 gives the brief overview of the existing work. Further, section 3 provides the details of the given tasks and dataset statistics. This is followed by the description of model used for experimentation in Section 4. The results are explained in the Section 5.

2. Background study

Several studies on sentiment analysis and the moderation of homophobic content on social media networks have been conducted; however, the majority have focused on high-resource languages such as English. To organise the related work, we divided the background study into two parts: section 2.1 provides a brief description of the model proposed for sentiment analysis, and part 2.2 describes the model proposed for homophobic content moderation.

2.1. Models proposed for Sentiment analysis

[10] developed a novel framework for assessing the rating of internet reviews. The suggested method detects polarity in online reviews by combining text processing and feature extraction methods. The authors claim that their proposed strategy outperforms existing deep learning methods. [11] used code-mixed text data from social media to identify sentiment. Their study made use of two code-mixed datasets: English-Bengali and English-Hindi. They grouped the data based on the statement's polarity conflict, such as positive, negative, or neutral. The translation and transliteration-based transformer model was developed by [12] to detect hateful comments from social media networks [13, 14, 15]. [16] presented a novel Framework for predicting discrepancies in Google App text comments and ratings using Deep Learning approaches. The

Table 1
Train and validation Kannada dataset

Category	Training	validation
Positive	2823	321
Negative	1188	139
Not-Kannada	916	110
Mixed feeling	574	52
Unknown state	711	69
Total	6212	691

framework is divided into two phases. In the first step, the polarity of reviews is predicted using a sentiment analysis algorithm. In the second step, star ratings are predicted from the text format of reviews after deep learning models have been trained on the ground truth obtained in the first phase.

2.2. Models proposed for Homophobic content detection

To extract homophobic information from social media data, [17] first convert code-mixed text to monolingual, utilising a data augmentation and transliteration-based approach. [18] used transformer-based XLM-Roberta to identify homophobia and transphobia data.TF-IDF vectorizer combined with SVM model is used by [19] to identify homophobia content. The number of monolingual and multilingual transformer models were experimented with data augmentation by [20] for homophobia detection.

3. Task and data description

DravidianLangTech organised the shared task on sentiment analysis and homophobia identification in YouTube comments [9][21]; The shared task included two different sub-tasks: Task A is Sentiment Analysis in Kannada, Malayalam, and Tamil, where we participated in the Kannada dataset, Task B is the detection of homophobic texts in English, Tamil, Tamil English, and Malayalam. The aim of sentiment analysis was to classify the code mixed data into positive, negative, and mixed feelings and not in the intended language. Classifying the code-mixed material into homophobic, transphobic, and non-anti-LGBTQ+ content was the goal of the second assignment.

The datasets for the competitions were made available in phases. Task A and task B training and validation datasets were released initially; later, Test data was made available. The dataset is collected from comments on popular YouTube channels. The dataset contains two fields: Text and Label. The complete statistics of the data we investigated in our work are presented in Table 1,2.

Table 2
Train and validation dataset for Homophobia detection

Dataset/category	Non anti LGBTQ+ Content	Homophobic	Transphobic	Total
Tamil Train	2022	485	155	2662
Tamil val	526	103	34	663
Malayalam Train	2434	491	189	3112
Malayalam val	692	133	41	866
English Train	3001	157	6	3160
English val	732	58	2	792
Tamil-Eng Train	3438	311	112	3861
Tamil-Eng val	862	66	38	966

4. Methodology

The current paper used the multi-class classification approach for sentiment analysis and homophobic and transphobic text detection. Several conventional machine learning models, and ensemble methods were used to realise the goal. A detailed description of all the methods is presented in the subsection below.

4.1. Data cleaning and pre-processing

The datasets were preprocessed before being fed into the models. The preprocessing is carried out on the Text field. The numbers, punctuation, and symbols have been deleted from the text because they do not help us predict the label. We also deleted white spaces; finally, the lower casing of text is performed to avoid redundant data. The cleaned texts are then tokenized and encoded into a series of token indexes.All of this preprocessing was done with the help of the NLTK toolbox from the Python library [1]. Furthermore, TF-IDF vectorization (n-gram vectors) is performed, and vectorized data is used as input for different models. We also applied SMOTE on vectorised data to balance the overall dataset.

4.2. Classification Models

We used different ensemble techniques, and traditional machine learning classifiers in the proposed approach to predict the outcomes. The following sections provide comprehensive details of each of these models.

4.2.1. Conventional Machine leaning classifiers

Initially, we experimented with different conventional machine learning models such as Logistics Regression, Passive Aggressive classifier, Support vector machine (SVM), Random Forest and Naïve Bayes to classify the text into their respective categories. We have used default parameters provided by the sci-kit-learn library to train the models. The input for all these models was

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

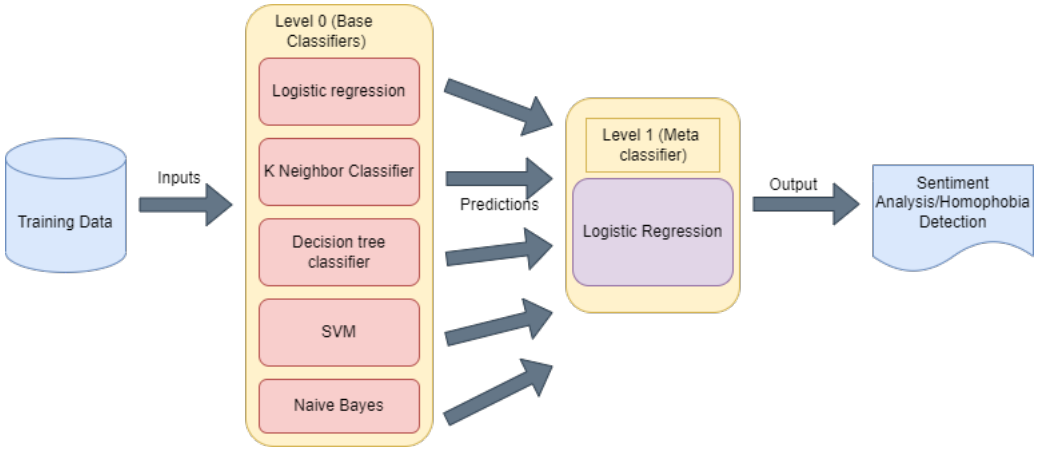

Figure 1: A stacking ensemble model

taken from TF-IDF vectors created from the cleaned text. The model was developed using Python's sci-kit-learn library [2].

4.2.2. Ensemble Machine Learning method

We employed an ensemble setup in the model to increase the performance of classic machine learning models. Three different ensemble approaches were used to classify the text: gradient ensemble, stacking ensemble, and model selection ensemble. As weak learners, the stacking ensemble included logistic regression,k nearest neighbour classifier, decision tree classifier, Support vector Machine (SVM), and naive Bayes classifier. The logistic regression, random forest classifier, and SVM were employed in the model selection and gradient boosting. The TF-IDF vectoriser is used as the input for all of these models. The detailed Architecture of the proposed model is illustrated in Fig 1.

5. Results

All experiments were conducted in the Keras and sklearn environments. To read the datasets, we utilised the pandas library. The dataset was prepared using Keras preprocessing methods and *nltk* library. Using sentiment and homophobic data provided by the task organisers, we used K-fold cross-validation to train our proposed models. Experimental trials are used to select the hyperparameter value K=5. Table 3 illustrates the findings of the sentiment analysis performed on the Kannada dataset, and Table 4 provides homophobia results.

For sentiment analysis using the Kannada dataset the best model was found to be the model using stacking ensemble with the accuracy of 0.515. The stacking ensemble consisted of Logistic Regression, KNeighbors Classifier, Decision Tree Classifier,SVM and Gaussian Naive Bayes as the base models and Logistic Regression as the meta learner model. Different models were used

[2]https://scikit-learn.org/stable/

Table 3
Models performance on Kannada sentiment validation dataset

Models	Score
Logistic Regression	0.496
Passive Aggressive	0.432
SVM	0.505
Naive Bayes	0.362
Random Forest	0.504
Gradient Boosting	0.494
Stacking Ensemble	0.515
Voting Ensemble	0.501

Table 4
Models performance on homophobia validation dataset

	Tamil	English	Malayalam	Tamil-English
Logistic Regression	0.760	0.930	0.812	0.747
Passive Aggressive	0.760	0.865	0.927	0.688
SVM	0.760	0.922	0.883	0.769
Naive Bayes	0.580	0.906	0.833	0.757
Gradient Boosting	0.759	0.916	0.825	0.891
Stacking Ensemble	0.762	0.978	0.925	0.890
Voting Ensemble	0.759	0.966	0.832	0.890

to detect homophobia for different datasets. The stacking ensemble produced the best results on the Tamil dataset, with an accuracy of 0.762. In stacking ensemble Logistic Regression, K nearest neighbours Classifier, Decision Tree Classifier, SVM, and Gaussian Naive Bayes were included as base learners, with Logistic Regression serving as the meta learner model. Similarly, the English dataset has given better results using the stacking ensemble model with an accuracy of 0.966. On the other hand, the Malayalam dataset performed best with the Passive Aggressive classifier, with an accuracy of 0.927. The model chosen for the Tamil English dataset was gradient boosting, which produced an accuracy of 0.891.

The organisers provided a weighted F1 score to evaluate the presented models. Our top-performing Stacking ensemble model was ranked 11th and 6th among the participating teams on Kannada, Tamil, and English datasets. Similarly, Passive Aggressive and gradient boosting performed better on Malayalam and Tanglish data, ranking second and fourth, respectively. Table 5 illustrates the final ranking of our proposed models among the participating teams.It also includes the best F1 scores achieved among the participating teams.

6. Conclusion and Future work

In our work, we presented a model submitted by our team for Sentiment analysis and Homophobia content identification on You Tube comments in the Fire 2022 shared task. Our proposed

Table 5
F1 score and ranks of the test dataset of Task A and Task B

	Model	F1 score	Rank	Best F1 Score
Kannada	Stacking Ensemble	0.35	11	0.550
Tamil	Stacking Ensemble	0.26	6	0.366
English	Stacking Ensemble	0.322	6	0.493
Malayalam	Passive Aggressive	0.94	2	0.974
Tamil English	Gradient Boosting	0.34	4	0.580

work evaluated two distinct models: a machine learning-based model and an ensemble setup with machine learning classifiers as base learners. The experimental findings demonstrate that ensemble models outperform different baseline models for stance detection. We can increase the efficiency of the suggested modes by using context-aware domain-specific embeddings.

References

[1] W. Medhat, A. Hassan, H. Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams engineering journal 5 (2014) 1093–1113.

[2] S. Saumya, J. P. Singh, Detection of spam reviews: a sentiment analysis approach, CSI Transactions on ICT 6 (2018) 137–148.

[3] U. Makhmudah, S. Bukhori, J. A. Putra, B. A. B. Yudha, Sentiment analysis of indonesian homosexual tweets using support vector machine method, in: 2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE), IEEE, 2019, pp. 183–186.

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

[5] A. M. Ramadhani, H. S. Goo, Twitter sentiment analysis using deep learning methods, in: 2017 7th International annual engineering seminar (InAES), IEEE, 2017, pp. 1–4.

[6] S. Biradar, S. Saumya, A. Chauhan, Combating the infodemic: Covid-19 induced fake news recognition in social media networks, Complex & Intelligent Systems (2022) 1–13.

[7] S. Biradar, S. Saumya, Iiitdwd@ tamilnlp-acl2022: Transformer-based approach to classify abusive content in dravidian code-mixed text, in: Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, 2022, pp. 100–104.

[8] R. Priyadharshini, B. R. Chakravarthi, S. Thavareesan, D. Chinnappa, D. Thenmozhi, R. Ponnusamy, Overview of the DravidianCodeMix 2021 shared task on sentiment detection in Tamil, Malayalam, and Kannada, in: Forum for Information Retrieval Evaluation, 2021, pp. 4–6.

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

[10] G. S. Budhi, R. Chiong, I. Pranata, Z. Hu, Using machine learning to predict the sentiment of online reviews: a new framework for comparative analysis, Archives of Computational Methods in Engineering 28 (2021) 2543–2566.
[11] S. Ghosh, S. Ghosh, D. Das, Sentiment identification in code-mixed social media text, arXiv preprint arXiv:1707.01184 (2017). doi:https://doi.org/10.48550/arXiv.1707.01184.
[12] S. Biradar, S. Saumya, et al., Fighting hate speech from bilingual hinglish speaker's perspective, a transformer-and translation-based approach., Social Network Analysis and Mining 12 (2022) 1–10.
[13] S. Saumya, A. Kumar, J. P. Singh, Offensive language identification in dravidian code mixed social media text, in: Proceedings of the first workshop on speech and language technologies for Dravidian languages, 2021, pp. 36–45.
[14] A. K. Mishra, S. Saumya, A. Kumar, Iiit_dwd@ hasoc 2020: Identifying offensive content in indo-european languages., in: FIRE (Working Notes), 2020, pp. 139–144.
[15] A. Kumar, S. Saumya, J. P. Singh, Nitp-ai-nlp@ hasoc-fire2020: Fine tuned bert for the hate speech and offensive content identification from social media., in: FIRE (Working Notes), 2020, pp. 266–273.
[16] S. Sadiq, M. Umer, S. Ullah, S. Mirjalili, V. Rupapara, M. Nappi, Discrepancy detection between actual user reviews and numeric ratings of google app store using deep learning, Expert Systems with Applications 181 (2021) 115111.
[17] B. R. Chakravarthi, A. Hande, R. Ponnusamy, P. K. Kumaresan, R. Priyadharshini, 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 (2022) 100119.
[18] J. García-Díaz, C. Caparrós-Laiz, R. Valencia-García, Umuteam@ lt-edi-acl2022: Detecting homophobic and transphobic comments in tamil, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, 2022, pp. 140–144.
[19] N. Ashraf, M. Taha, A. Abd Elfattah, H. Nayel, 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, 2022, pp. 287–290.
[20] V. Bhandari, P. Goyal, bitsa_nlp@lt-edi-acl2022: Leveraging pretrained language models for detecting homophobia and transphobia in social media comments, arXiv preprint arXiv:2203.14267 (2022).
[21] K. Shumugavadivel, M. Subramanian, P. K. Kumaresan, B. R. Chakravarthi, B. B, S. Chinnaudayar Navaneethakrishnan, L. S.K, T. Mandl, R. Ponnusamy, V. Palanikumar, M. Balaji J, Overview of the Shared Task on Sentiment Analysis and Homophobia Detection of YouTube Comments in Code-Mixed Dravidian Languages, in: Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, CEUR, 2022.