NAYEL @LT-EDI-ACL2022: Homophobia/Transphobia Detection for Equality, Diversity, and Inclusion using SVM

Nsrin Ashraf and **Mohamed Taha** and **Ahmed Taha** and **Hamada Nayel**
Department of Computer Science, Benha University, Benha, Egypt
{nisrien.ashraf19,mohamed.taha}@fci.bu.edu.eg
{ahmed.taha, hamada.ali}@fci.bu.edu.eg

Abstract

Analysing the contents of social media platforms such as YouTube, Facebook and Twitter gained interest due to the vast number of users. One of the important tasks is homophobia/transphobia detection. This paper illustrates the system submitted by our team for the homophobia/transphobia detection in social media comments shared task. A machine learning-based model has been designed and various classification algorithms have been implemented for automatic detection of homophobia in YouTube comments. TF-IDF has been used with a range of bigram models for vectorization of comments. Support Vector Machines have been used to develop the proposed model and our submission reported 0.91, 0.92, 0.88 weighted F1-scores for English, Tamil and Tamil-English datasets respectively.

1 Introduction

The internet provides a wealth of information that is immensely useful for different reasons. Due to the overwhelming information available on the internet, online social media platforms inspired a new epoch of "misinformation" by spreading incorrect or misleading information to delude users. Social media platforms such as YouTube, Facebook, Twitter, and other platforms initially became popular due to the social aspects that they allow users to post and share material to share their opinions and ideas on anything at any time. YouTube is a popular platform that allows users to create their accounts, upload videos, and make comments. Due to the massive audience, distributing negative or uncomfortable information has become easier. There is a need for developing tools for automatic detection of different behaviours such as fake news, sentiment analysis, hate speech, aggressive content and rumours. YouTube is one of the most popular social media platforms, in which any user can share data about anything without any restrictions that data may contain scandalous data such as racist, homophobic, transphobic, and antiLGBT+ propaganda.(Jagtap et al., 2021)

Since misinformation spreads faster than factual news among people, we need to classify this information whether containing LGBT+ data or not. In the proposed system the dataset used is collected from YouTube comments and divided into three datasets with various languages namely; English, Tamil and mixed languages Tamil-English. The proposed model uses a machine learning approach integrated with text vectorization to develop a system for automatic detection of Homophobic or Transphobic contents.

2 Background

Pathak et al. (2021) developed a machine learning based model for hate speech and offensive language detection. They used a multilingual dataset consisting of tweets and YouTube comments written in Malayalam, Tamil and English. TF-IDF and word embeddings were used for feature extraction phase. They trained different machine learning classifiers and they used 5-fold cross-validation approach to evaluate the performance of the classifiers. Multinomial Naive Bayes (MNB) reported the best F1-score 77% for Malayalam-English dataset, while Support Vector Machines (SVMs) obtained the best F1-score 87% for Tamil-English dataset. Nayel (2020) used TF-IDF as weighting scheme with a range of n-gram for feature extraction to implement Stochastic Gradient Descent (SGD) algorithm for automatic offensive language detection in Arabic tweets. Nayel and L (2019) developed a model for Hate Speech detection in multilingual contents using SVM and Multi-Layer Perceptron (MLP). TF-IDF model as vector representation of collected tweets. SVM reported the best F1-score

Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 287 - 290

for English dataset, while MLP reported the best F1-score for German and Hindi languages.

Though much work has been done to identify offensive content in major languages such as English (Chakravarthi et al., 2021), it is a challenging task to identify and flag offensive content in low-resource languages because many users prefer to write their language in English script, a practise known as code-switching or code-mixing (Hande et al., 2021; Nayel et al., 2021).

3 Dataset

The dataset given for the shared task (Chakravarthi et al., 2022) consists of YouTube comments in three languages English, Tamil and the remaining code-mixed Tamil-English. The dataset contains some unique features that distinguish it from prior hate speech or offensive language identification datasets The extracted comments including Homophobic and Non-anti-LGBT+ text. These comments have been scraped using a scraper tool and were collected between August 2020 and Feb 2021. Table 1 shows the statistics of the tweets, indicating that the total number of comments in three languages which is about 22K . The full details of the dataset is given in (Chakravarthi et al., 2021).

4 System Overview

In this section, we review the phases of the proposed model. The primary aim of this work is to explore the impact of different machine learning methods on automatic Homophobia/Transphobia detection in social media comments. The proposed model composite of the following phases:

4.1 Text Cleaning

In this phase, some basic preprocessing steps have been carried out. The aim of this step is to clean the raw text from unwanted information. These steps includes:-

- Hashtag and special symbols removal,
- URL and whitespace removal,
- Repeated character removal.

4.2 Features Engineering

Extracting the features from comments is an essential step for building the classification model. This comes directly after preprocessing step. In this work Term Frequency/Inverse Document Frequency (TF-IDF) technique was used as vector space model that represents the comments as vector of real numbers.

4.3 Methods

Different classification algorithms have been implemented as well as ensemble approach using hard voting. TF-IDF has been used as a vector space model for comments representation. The set of classification algorithms that have been used are listed bellow.

1. **Support Vector Machine (SVM)**
   As we are classifying text based on a wide feature set for a binary classification problem and is available in various kernels function. The objective of SVM algorithm is to estimate a hyperplane based on feature set to classify data points (Nayel, 2019).

2. **Random Forest (RF)**
   RF is an advanced form of decision trees which is a supervised learning model. RF consists of many decision trees working individually to estimate the result of a class, with the final predictions based on the class with the most votes (Breiman, 2001).

3. **Passive Aggressive Classifier (PA)**
   It has shown to be a very successful and popular way for online learning to address many real-world issues (Crammer et al., 2006). Online learning is utilized in circumstances where there is a requirement to keep a regular check on the data, such as news, social media, and so on. The main premise of this algorithm is that it examines data, learns from it, and discards it without keeping it. When there is a misclassification, the algorithm responds aggressively by changing the values, and when there is a right classification, it responds lazily or passively.

4. **Gaussian Naïve Bayes (GNB)**
   It is a supervised learning classifier based on the Bayes theorem that calculates explicit probabilities for hypotheses and provides a useful perspective for understanding many learning algorithms that do not explicitly manipulate probabilities (Ontivero-Ortega et al.,

Language	Number of comments	Number of tokens	Number of charaters
English	7,265	116,015	632,221
Tamil	5,240	255,578	787,177
Tamil-English	10,319	88,303	628,077
Total	22,824	249,896	2,047,475

Table 1: Raw dataset statistics by language

2017). Gaussian Naïve Bayes is the most important among the categories of Naïve Bayes because the classifier is used when the predictor values are continuous and are expected to follow a Gaussian distribution.

5. **Multi-Layer Perceptron (MLP)**
MLP is a feed-forward neural network augmentation. It is made up of three layers: the input layer, the output layer, and the hidden layer (Hopfield, 1988). The input signal to be processed is received by the input layer. The output layer is responsible for tasks such as prediction and categorization. The real computational engine of the MLP is an arbitrary number of hidden layers inserted between the input and output layers. In an MLP, data flow in the forward direction from input to output layer, like a feed-forward network. The back-propagation learning technique is used to train the neurons in the MLP. MLPs are intended to approximate any continuous function and can solve problems that are not linearly separable.

5 Experimental Setup

- F1-score has been used to evaluate the performance of all submissions. F1-score is the harmonic mean of Precision (P) and Recall (R) and calculated as follows:

$$F\text{–}score = \frac{2 * P * R}{P + R}$$

- Bi-gram model have been used while calculating TF-IDF for the the entire dataset.
- For validation purpose, cross-validation technique has been used and the training set has been divided into five folds.
- For SVM, linear kernel has been tested with regularization parameter set to 5.
- The number of nodes in the hidden layer of MLP was set at 20, logistic function was used as activation function and Adam solver was used with maximum number of iterations set to 200. The maximum number of decision trees in random forests is set at 300.

Table 2: 5-fold cross validation F1-score for development phase

Classifier	English	Tamil	Tamil-English
SVM	0.43	0.85	0.51
RF	0.34	0.85	0.35
PA	0.42	0.85	0.51
GNB	0.40	0.70	0.41
MLP	0.37	0.84	0.47

6 Results and Discussion

Table 2 shows the F1-score reported at development phase for different classifiers with different language. It is clear that, Tamil dataset reported the best performance while English dataset reported the worst. SVM for all datasets outperformed all other classifiers.

Table 3 shows the final results of SVM for all datasets. It is clear that weighted F1-score (W F1-score) reports high values for all datasets, while macro F1-score (M F1-score) reported lowest values. The results shown in Table 3 show that the performance of SVM achieved better results on Tamil dataset. Our model for Tamil achieved the second rank, while in English our model achieved 11th rank. The proposed model for mixed-code dataset achieved 8th rank.

7 Conclusion

In this work we implemented a machine learning model using SVM as a classification algorithm for homophobia/transphobia detection in text. The comments have been represented as TF-IDF vectors.

Table 3: Detailed results of SVM for test set on all languages

Dataset	English	Tamil	Tamil-English
Accuracy	0.94	0.92	0.90
M F1-score	0.39	0.84	0.51
W F1-score	0.91	0.92	0.88
Rank	11	2	8

Applying more complex systems may improve the performance of the model. Deep learning based models have various structure that can enhance the output. Another word representation models such as word embeddings can be used as input for better representation.

References

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

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

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kayalvizhi Sampath, Durairaj Thenmozhi, Sathiyaraj Thangasamy, Rajendran Nallathambi, and John Phillip McCrae. 2021. Dataset for identification of homophobia and transophobia in multilingual youtube comments. *arXiv preprint arXiv:2109.00227*.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7(19):551–585.

Adeep Hande, Karthik Puranik, Konthala Yasaswini, Ruba Priyadharshini, Sajeetha Thavareesan, Anbukkarasi Sampath, Kogilavani Shanmugavadivel, Durairaj Thenmozhi, and Bharathi Raja Chakravarthi. 2021. Offensive language identification in low-resourced code-mixed dravidian languages using pseudo-labeling. *arXiv preprint arXiv:2108.12177*.

John J Hopfield. 1988. Artificial neural networks. *IEEE Circuits and Devices Magazine*, 4(5):3–10.

Raj Jagtap, Abhinav Kumar, Rahul Goel, Shakshi Sharma, Rajesh Sharma, and Clint P. George. 2021. Misinformation detection on youtube using video captions. *CoRR*, abs/2107.00941.

Hamada Nayel. 2020. NAYEL at SemEval-2020 task 12: TF/IDF-based approach for automatic offensive language detection in Arabic tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2086–2089, Barcelona (online). International Committee for Computational Linguistics.

Hamada Nayel, Eslam Amer, Aya Allam, and Hanya Abdallah. 2021. Machine learning-based model for sentiment and sarcasm detection. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 386–389, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Hamada A. Nayel. 2019. NAYEL@APDA: Machine Learning Approach for Author Profiling and Deception Detection in Arabic Texts. In *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019*, volume 2517 of *CEUR Workshop Proceedings*, pages 92–99. CEUR-WS.org.

Hamada A. Nayel and Shashirekha H. L. 2019. DEEP at HASOC2019: A machine learning framework for hate speech and offensive language detection. In *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019*, volume 2517 of *CEUR Workshop Proceedings*, pages 336–343. CEUR-WS.org.

Marlis Ontivero-Ortega, Agustin Lage-Castellanos, Giancarlo Valente, Rainer Goebel, and Mitchell Valdes-Sosa. 2017. Fast gaussian naïve bayes for searchlight classification analysis. *NeuroImage*, 163:471–479.

Varsha M. Pathak, Manish Joshi, Prasad Joshi, Monica Mundada, and Tanmay Joshi. 2021. Kbcnmujal@hasoc-dravidian-codemix-fire2020: Using machine learning for detection of hate speech and offensive code-mixed social media text. *CoRR*, abs/2102.09866.