Homophobia, Transphobia Detection in Tamil, Malayalam, English Languages using Logistic Regression and Code-Mixed Data using AWD-LSTM ⋆

SIVAPRASATH S[1,*,†], Lavanya Sambath Kumar[2,†] and Sajeetha Thavareesan[3,†]

[1]SRM Institute of Science and Technology, Chennai, India.
[2]SRM Institute of Science and Technology, Chennai, India.
[3]Eastern University, Sri Lanka.

Abstract

This paper presents the submission of the shared task "Homophobia, Transphobia Detection of YouTube Comments" organized by DravidianLangTech. Our team has participated in Task - B, which tries to identify the comments on youtube are Non-anti LGBTQ+ content or Homophobic or Transphobic in code-mixed(Tamil-English), Tamil, Malayalam, and English. We proposed the AWD-LSTM model for the code-mixed(Tamil-English) language data set and Logistic Regression for Tamil, Malayalam, and English languages. Our AWD-LSTM model achieved a 0.33 macro average F1 score for code-mixed(Tamil-English) language and Logistic Regression achieved a 0.55 macro average F1 score in the Tamil language, 0.98 macro average F1-score in the Malayalam language, 0.91 macro average F1-score in the English language.

Keywords

Homophobia, code-mix, AWD-LSTM, YouTube, Logistic Regression

1. Introduction

Internet users have access to a vast amount of information that can be of great benefit to them for a variety of reasons. Online social media platforms have been credited with igniting a new phase of "misinformation" by sharing wrong or incorrect information with the intention of misleading users. On the well-known application "YouTube", users can create their own accounts, upload videos, and leave comments [17]. In recent decades, as an increasing number of people have gained access to the internet, it has become highly crucial to track their activity in order to recognize behavior that is offensive towards any particular group based on sex, or sexual orientation to maintain the Internet's inclusivity, support diversity in ideas and content, and promote innovation. This task has 4 sub-tasks. First is Homophobia, Transphobia detection in English Language, Second in Malayalam Language, third in Tamil Language and fourth in code-mixed(Tamil-English) language. At first, English was the only language that could be used for communication. We are fortunate to be able to communicate with one another in any language. Because of this, many individuals now speak a hybrid language that combines English

FIRE 2022 - Forum for Information Retrieval Evaluation, December 9-13, 2022, India.
*Corresponding author.
†These authors contributed equally.
ss4483@srmist.edu.in (S. S); lavanyas6@srmist.edu.in (L. S. Kumar); sajeethas@esn.ac.lk (S. Thavareesan)

with their own original language, also known as their mother tongue [13, 18, 19]. The practice of changing between two or more different languages during communication is known as code-mixing. The most prevalent language is Hindi-English, but this challenge included code-mixed texts Tamil and English [15]. Tamil and Malayalam are members of the Dravidian language family. One of the most widely used social media sites is YouTube, where anyone may publish information about anything without limits. Transphobic and homophobic content indicates statements expressing hatred for lesbian, gay, transgender, and bisexual individuals. It results in offensive speech and generates major societal difficulties that can make online platforms toxic and unpleasant for LGBT+ individuals. This information may contain controversial information against LGBTQ+ members also. So, in order to prevent this happen again, we need to classify the text and delete the comments which are transphobic and homophobic automatically. We must categorize this data whether it is non-LGBTQ+ content or homophobic or Transphobic since false information spreads among individuals more quickly than accurate news [20, 21].

2. Related works

T.Mandl et.al has created a model for analyzing hate speech in German, Hindi, and English. They employed a model called long-term short-term memory. They achieved a macro - F1 score of 0.78 for English, 0.81 for Hindi, and 0.61 for German [12]. TF-IDF-oriented multi-view SVM as well as keyword-based approaches. This method is superior to other neural methods in its effectiveness. It employs multiple view stacked SVM(Support Vector Machines). This model got 0.80 accuracy in the Stormfront data [11].

In their presentation on the detection of offensive comments in the Manglish data set, Sara Renjit and Sumam Mary Idicula use Keras embedding layers with LSTM layers. Offensive and non-offensive comments were classified based on the embedding model. They have got 0.53 weighed average F1 score based on Keras embeddings and 0.48 weighed average F1 score based on Doc2Vec [13]. A corpus for performing sentiment analysis using the Manglish language was provided by B. R. Chakravarthi and colleagues. For Manglish(Code-mixed), they presented a gold standard corpus. This corpus has got Krippendorff's alpha of more than 0.8. Volunteers were responsible for annotating the data set [3].

A data set with Hindi and English(code-mixed) was provided by Prabhu et.al. They have implemented LSTM architecture representation at the sub-word level. They have collected comments from Facebook pages that are publicly available and created their dataset. They achieved an F1-score of 0.658 in their dataset and an F1-score of 0.537 in the SemEval'13 data set. They were able to attain an accuracy that was 4-5% higher than other methods [9]. Nsrin Ashraf et.al employed an SVM model to detect homophobia and transphobia, and they used TF-IDF to vectorize the data. They obtained weighed F1-scores and accuracy of 0.92 and 0.92 for Tamil respectively, 0.88 and 0.90 for Tanglish(code-mixed) respectively, and 0.91 and 0.94 for English respectively [2].

3. Data set description

The data set for the YouTube comments in Tamil, Malayalam, English, Code - Mixed(Tamil-English) has been given by the organizer. There are three labels in each data set. External data sets are not used, and labels are not given for any of the test data sets.

The labels given in these data sets are,

- Non-anti LGBTQ+ content
- Homophobic
- Transphobic

4. Text cleaning

Some fundamental pre-processing operations have been done in this phase. The purpose of this phase is to remove unnecessary information from the raw text. All the stop words, URLs, symbols $,? @, =, etc, and emojis were removed.

5. Proposed methods

For the code-mixed(Tamil-English) data set AWD-LSTM model is used. It is one of the most used language models. It is used to predict and analyze the sequence, especially in natural language processing. AWD-LSTM uses DropConnect, along with a few other well-known regularisation algorithms, including a variation of Average-SGD (NT-ASGD) and average random gradient descent method, are used. It introduced the variant SGD—ASGD. The ASGD algorithm employs the same gradient update step as the SGD technique, but it delivers an average value rather than the weight determined in the most recent iteration.

For Tamil, Malayalam, and English languages, the Logistic Regression algorithm is used. The logistic regression classifier builds a frequency dictionary mapping using a list of words as input. After that, the logistic regression classifier will be trained while the cost function is minimized. Following training, test data sets can be fed into the system to make predictions.

6. Implementation

All the necessary modules and packages like Numpy, pandas, TensorFlow, RegEx, scikit learn are imported.

6.1. Code-Mixed(Tam-Eng)

Data set is imported and it is pre-processed. In pre-processing, the upper cases were converted to lower cases and unwanted columns were removed. We have used fast text word embeddings. Using a code-mixed sentence piece tokenizer, all the text was tokenized. We have used a pre-trained AWD-LSTM model which is already trained on another dataset. There are seven layers,

they are BatchNorm 1d, Dropout, Linear, ReLU, BatchNorm 1d, Dropout, and Linear. The first layer batchNorm 1d with a number of features 1200, the second layer Dropout with rate 0.2, the third layer Linear with in_features 1200, out_features 50, the fourth layer ReLU, the fifth layer BatchNorm 1d with a number of features 50, sixth layer Dropout with rate 0.1, seventh layer Linear with in_features 50, out_features 3.

6.2. Tamil, Malayalam and English

All the packages and modules were imported. The data set is imported and emojis and stop words from the text were removed. All the special characters were removed and uppercase letters were converted to lowercase letters. All the input texts were tokenized. Tfidf Vectorizer was used for vectorizing the text. The Logistic Regression model was trained with class_weight='balanced', C =5, and the model used for prediction.

7. Results and Discussion

7.1. Code-Mixed(Tam-Eng)

The metrics accuracy, precision, recall, and macro average F1-score have been calculated for Code-mixed(Tam-Eng) languages. The accuracy of 0.87, macro precision of 0.35, recall of 0.34, and macro average F1-score of 0.33 were recorded for Code-mixed(Tami-Eng) language. Due to class imbalance in the data set, this model performs comparatively poorly. Macro average F1 scores of other teams with our team are shown in Table 1 for code-mixed(Tam-Eng) language.

Table 1
Macro F1- score for code-mixed(Tam-Eng) Language of other teams with our team "SENTIZEN"

S.No	Team	Macro - F1
1	mucs	0.580
2	fnet	0.555
3	CITK	0.477
4	nlpzip	0.393
5	qwerty	0.344
6	IRLab@IITBHU	0.333
7	SSN-CSE-2022	0.316
8	BharataNLP	0.316
9	**SENTIZEN**	**0.330**

7.2. Tamil, Malayalam and English

The metrics accuracy, macro average precision, macro average recall, and macro average F1-score have been calculated for Tamil, Malayalam, and English languages.

Macro average F1 scores of other teams with our team are shown in Table 2 for the Tamil language.

Table 2
Macro F1- score for the Tamil Language of other teams with our team "SENTIZEN"

S.No	Team	Macro - F1
1	mucs	0.366
2	fnet	0.327
3	CITK	0.290
4	IRLab@IITBHU	0.289
5	qwerty	0.234
6	SSN-CSE-2022	0.234
7	BharataNLP	0.234
8	nlpzip	0.228
9	**SENTIZEN**	**0.310**

The accuracy of 0.55, macro average precision of 0.37, macro average recall of 0.49, macro average F1-score of 0.31 were recorded for the Tamil language. Our model performed comparatively better than other teams for the Tamil Language.

Macro average F1 scores of other teams with our team are shown in Table 3 for the Malayalam language.

Table 3
Macro F1- score for the Malayalam Language of other teams with our team "SENTIZEN"

S.No	Team	Macro - F1
1	Nltk	0.974
2	qwerty	0.943
3	BharataNLP	0.942
4	CITK	0.860
5	mucs	0.750
6	fnet	0.696
7	nlpzip	0.542
8	IRLab@IITBHU	0.427
9	SSN-CSE-2022	0.296
10	**SENTIZEN**	**0.970**

The accuracy of 0.98, macro average precision of 0.98, macro average recall of 0.96, and F1-score of 0.97 were recorded for the Malayalam language. Our model's performance is better than other teams for the Malayalam language.

Macro average F1 scores of other teams with our team are shown in Table 4 for the English language.

Table 4
Macro F1- score for the English Language of other teams with our team "SENTIZEN"

S.No	Team	Macro - F1
1	BharataNLP	0.493
2	fnet	0.486
3	nlpzip	0.462
4	mucs	0.374
5	IRLab@IITBHU	0.337
6	qwerty	0.332
7	SSN-CSE-2022	0.332
8	kongu.eng-21MSR002	0.319
9	**SENTIZEN**	**0.420**

The accuracy of 0.91, macro average precision of 0.42, macro average recall of 0.42, and F1-score of 0.42 were recorded for the English language. Our model performed comparatively better than other teams for the English Language.

8. Conclusion

This shared task aims at detecting homophobia, and transphobia detection of YouTube comments on Tamil, Malayalam, English, and Code-mixed languages. There were total 9,9,10 and 10 teams that participated in code-mixed, Tamil, Malayalam, and English languages respectively. We have used the AWD-LSTM model for code-mixed(Tam-Eng) language and achieved an accuracy of 0.87. For Tamil, Malayalam, and English languages we used Logistic regression and achieved an accuracy of 0.55, 0.98, and 0.91 respectively. Our team's macro average F1 score for the Malayalam language is better compared with other teams. In the future, we will consider improving the model's performance by class balancing in the given data.

References

[1] Arora, G. (2020). Gauravaroin@ HASOC-Dravidian-CodeMix-FIRE2020: pre-training ULMFiT on synthetically generated code-mixed data for hate speech detection. arXiv preprint arXiv:2010.02094.

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

[3] Chakravarthi, B. R., Jose, N., Suryawanshi, S., Sherly, E., & McCrae, J. P. (2020). A sentiment analysis dataset for code-mixed Malayalam-English. arXiv preprint arXiv:2006.00210.

[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] Dahiya, A., Battan, N., Shrivastava, M., & Sharma, D. M. (2019, August). Curriculum Learning Strategies for Hindi-English Code-Mixed Sentiment Analysis. In International Joint Conference on Artificial Intelligence (pp. 177-189). Springer, Cham.

[6] Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR), 51(4), 1-30.

[7] Gupta, D., Tripathi, S., Ekbal, A., & Bhattacharyya, P. (2017). SMPOST: parts of speech tagger for code-mixed indic social media text. arXiv preprint arXiv:1702.00167.

[8] Jose, N., Chakravarthi, B. R., Suryawanshi, S., Sherly, E., & McCrae, J. P. (2020, March). A survey of current datasets for code-switching research. In 2020 6th international conference on advanced computing and communication systems (ICACCS) (pp. 136-141). IEEE.

[9] Joshi, A., Prabhu, A., Shrivastava, M., & Varma, V. (2016, December). Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (pp. 2482-2491).

[10] Kedia, K., & Nandy, A. (2021). indicnlp@ kgp at DravidianLangTech-EACL2021: Offensive language identification in Dravidian languages. arXiv preprint arXiv:2102.07150.

[11] MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. PloS one, 14(8), e0221152.

[12] Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., & Patel, A. (2019, December). Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In Proceedings of the 11th forum for information retrieval evaluation (pp. 14-17).

[13] Renjit, S., & Idicula, S. M. (2020). CUSATNLP@ HASOC-Dravidian-CodeMix-FIRE2020: Identifying Offensive Language from ManglishTweets. arXiv preprint arXiv:2010.08756.

[14] Shumugavadivel, Kogilavani and Subramanian, Malliga and Kumaresan, Prasanna Kumar and Chakravarthi, Bharathi Raja and B, Bharathi and Chinnaudayar Navaneethakrishnan, Subalalitha and S.K, Lavanya and Mandl, Thomas and Ponnusamy, Rahul and Palanikumar, Vasanth and Balaji J, Manoj. Overview of the Shared Task on Sentiment Analysis and Homophobia Detection of YouTube Comments in Code-Mixed Dravidian Languages. In proceedings of dravidiancodemix-2022.

[15] Zhang, Y., Riesa, J., Gillick, D., Bakalov, A., Baldridge, J., & Weiss, D. (2018). A fast, compact, accurate model for language identification of codemixed text. arXiv preprint arXiv:1810.04142.

[16] Zhu, Y., & Zhou, X. (2020). Zyy1510@ HASOC-Dravidian-CodeMix-FIRE2020: An Ensemble Model for Offensive Language Identification. In FIRE (Working Notes) (pp. 397-403).

[17] Subramanian, M., Ponnusamy, R., Benhur, S., Shanmugavadivel, K., Ganesan, A., Ravi, D., Shanmugasundaram, G.K., Priyadharshini, R. and Chakravarthi, B.R., 2022. Offensive language detection in Tamil YouTube comments by adapters and cross-domain knowledge transfer. Computer Speech & Language, 76, p.101404.

[18] Shanmugavadivel, K., Sampath, S.H., Nandhakumar, P., Mahalingam, P., Subramanian, M., Kumaresan, P.K. and Priyadharshini, R., 2022. An analysis of machine learning models for sentiment analysis of Tamil code-mixed data. Computer Speech & Language, p.101407.

[19] Chakravarthi, B.R., Hande, A., Ponnusamy, R., Kumaresan, P.K. and 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), p.100119.

[20] Chakravarthi, B.R., 2022. Hope speech detection in YouTube comments. Social Network Analysis and Mining, 12(1), pp.1-19.

[21] Chakravarthi, B.R., 2022. Multilingual hope speech detection in English and Dravidian languages. International Journal of Data Science and Analytics, 14(4), pp.389-406.