Homophobia and Transphobia Detection of Youtube Comments in Code-Mixed Dravidian Languages using Deep learning

P Pranith, V Samhita, D Sarath and Durairaj Thenmozhi

Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam, Chennai

Abstract

Homophobia and Transphobia Detection is the task of identifying homophobia, transphobia, and non-anti-LGBT+ content from the given corpus. Homophobia and transphobia are both toxic languages directed at LGBTQ+(Lesbian, Gay, Bisexual, Transgender, and Queer community) individuals that are described as hate speech. The shared task we worked on was to try and predict if a given comment was homophobic in nature. We were provided with a corpus of comments in Tamil, Malayalam, Tamil-English, and English. We used the IndicBERT and LaBSE machine learning models to predict the content. The results were as follows: IndicBERT was used to train Tamil, Malayalam, and Tamil-English and LaBSE was used to predict the content in English. We achieved weighted average F1 scores of 0.46, 0.54, 0.39, and 0.28 for English, Malayalam, Tamil-English, and Tamil respectively.[1]

Keywords

Homophobia, Transphobia, LGBTQ+, IndicBERT, Transformers, LaBSE, Tokenizer

1. Introduction

With the advancing reach of social media in today's world, an increasing number of people are getting online to consume content, share messages and consequently express their views and opinions.[2] However, people often misuse this freedom to make comments propagating hatred and toxicity. YouTube, especially, is a popular platform due to the ease with which users can share content (videos, posts, shots) and like, share, and comment on said content. The downside to this is that it gives room for more online harassment and overt cyberbullying.[3] This often has a drastic impact on the lives of the affected individuals/communities. The LGBTQ+ community especially has often been subjected to such hate and bullying. Sexual orientation and gender identity[4] are imperative elements that constitute a person's identity and must hence, never be discriminated against. Yet, individuals of the community are victims of harmful and ignorant comments[5] , abuse and threats. The number of hate crimes against the community continues to be on the rise with the increasing use of social media.

The shared task focuses on trying to identify comments that are homophobic or transphobic and could potentially cause a lot of harm. Hence, we employ the use of deep learning-based NLP models that can be modeled to detect such homophobic and transphobic content to help

Forum for Information Retrieval Evaluation, December 09-13, 2022, India

pranith2010245@ssn.edu.in (P. Pranith); samhita2010017@ssn.edu.in (V. Samhita); sarath2010125@ssn.edu.in (D. Sarath); theni_d@ssn.edu.in (D. Thenmozhi)

Table 1
Dataset Description

Dataset	Non anti LGBTQ+ Content	Anti LGBTQ+ Content	Total
Tamil Train	2022	640	2662
Tamil val	526	137	663
Malayalam Train	2434	680	3112
Malayalam val	692	174	866
English Train	3001	163	3160
English val	732	60	792
Tamil-Eng Train	3438	423	3861
Tamil-Eng val	862	104	966

the social media platforms better deal with the same and ensure social media remains safe for all.

2. Task Description

Homophobia and Transphobia are toxic languages aimed at the LGBTQ+ community. The goal of shared task (task B) is to develop systems to identify such homophobic, transphobic, and non-anti-LGBT+ content from the given corpus[3] and thus accordingly predict the nature of subsequent data fed into the system to put into place measures to deal with hate towards the LGBTQ+ community. The task is pursued from our viewpoint as binary labelled classification: non-anti-LGBT+ ; homophobic/transphobic.

3. Dataset Description

The corpus[3] consisted of a collection of comments from Youtube and other social media platforms. The data consisted of four variants of the comments; namely Tamil, Malayalam, English, and Tamil-English. Each comment was provided a label indicating its nature (Homophobic, Transphobic, Non-anti-LGBT+)

The data was split into three sections; one to be used for training and one other for development. The Tamil, English, and Tamil-English files were provided in CSV format, while the Malayalam file was provided in TSV format.

The third and final section consisted of the test data which was provided in CSV format. The test data consisted of only the comment and the respective ID, no labels were provided. The test data was fed into the fully developed model to output the required results.

4. Methodology

The task was approached as a binary-label text classification task with the use of transformers. The data is preprocessed and then data is made suitable for training by creating a data frame

for all the instances of the data. Model training is done by fine-tuning the parameters of two transformers that include LaBSE[6], and IndicBERT[7].

4.1. Data Preparation

Data processing is important for any machine learning problem. The data available is often incoherent and unorganized and hence must be cleaned to bring it to a more organized format that can easily be processed and used to draw results from. The irregularities in the data are removed in the preprocessing stage and made suitable for training. Preprocessing steps for this machine learning problem involved different steps as follows:

- **Removal of special character** such as '#' , '@', ';' and '!' etc - Special characters such as '#,':' and '@' are very commonly used in comments in social media platforms such as YouTube. These characters can affect the training of the model as the words that end with special characters are treated differently by the model. Hence they must be removed before the training so as to not interfere with the training of the model.
- **Replacement of emojis** with the appropriate words : Emoticons convey emotional expression in a text and hence must be preserved. In order to take into account the emotions that they bring to a message, they are replaced with appropriate words that convey the same meaning.

The training and development data were concatenated and put into a data frame. The eighty-twenty rule was then used and eighty percent of the data was used as training data and twenty percent of the data was used to test the model's performance. Label encoding was done to the training dataset and testing dataset to convert them into machine-readable form. The subsequent f1 score, recall, and precision obtained were evaluated.

Sample input data:

Text	Label
உலக முடிவு என்பது.... நீதிபதியிடம் இருந்து.... ஆரம்பம் ஆகிறது.....	0
They harass everyone in bus and do this for living	0
Supper sister God bless you family	1
ഇന്ന് ആണ്കുട്ടികളും സുരക്ഷിതര് ഒന്നും അല്ലെടോ	0

Key: 0: Homophobic or Transphobic ; 1: Non-Anti LGBTQ+. Sample input data

4.2. Training the models

The transformer models namely IndicBERT and LaBSE were used whose implementation is explained below. Both the models were trained for 3 epochs. (the rest of the hyper parameters were restricted to their default values) The IndicBERT uses an autotokeniser with the accents set to true.

4.2.1. AI4Bharat/IndicBERT

IndicBERT[7] is a multilingual ALBERT[8] model pretrained in 12 of the most important Indian languages. It is pre-trained on the monolingual corpus of AI4Bharat, which contains about 9 billion tokens, and then assessed on a variety of tasks. IndicBERT works on par with or better than other multilingual models (mBERT, XLM-R, etc.) despite using fewer parameters.

The languages that IndicBERT covers are: Assamese, Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, Telugu.

For our implementation, the AI4Bharat/IndicBERT model was used on the vernacular languages (Tamil, Malayalam and Tamil-English)

4.2.2. LaBSE

LaBSE, a multilingual embedding model[6] is a powerful tool that may be used for a variety of downstream tasks, including text classification, clustering, and others. It does this by encoding text from several languages into a common embedding space. It accomplishes this while utilising semantic data to comprehend language. In order to promote consistency amongst the sentence embeddings, existing methods for creating these embeddings, such as LASER or mUSE, rely on parallel data, mapping a sentence from one language to another directly.

The LaBSE model was used for the English dataset.

4.3. Post processing

After training, the sample test data was generated using twenty percent of the concatenated training and development and fed into the model to evaluate its performance. Once the performance metrics were verified, the model was run on the actual test data provided.

5. Results

LaBSE model that was used on the English dataset achieved a mean F1-score of 0.4625 and our team was ranked third for the same. IndicBERT model was used for Tamil-English, Malayalam and Tamil datasets. It achieved a mean F1-score of 0.393 for the Tamil-English dataset and we were placed third. It achieved a mean F1-score of 0.542 for the Malayalam dataset and we were placed seventh. For the tamil dataset, the model achieved a mean F1-score of 0.228 and was ranked eighth. The following performance metrics[9] were used to evaluate the predicted labels.The required formulas are provided below:

Precision: Precision refers to the probability that a correct classification has been done. It is the ratio of true positives to total positives predicted.

$$Precision = \frac{TP}{TP + FP}$$

Recall: Recall may be defined as the number of positives returned by our ML model. Recall is the measure of the model correctly identifying the True Positives

$$Recall = \frac{TP}{TP + FN}$$

F1 score: The F1 score metric uses a combination of precision and recall to assess the performance of the ML model. The F1 score is harmonic mean of the two.

$$F1Score = 2.\frac{precision.recall}{precision + recall}$$

Accuracy: It is the number of correct predictions to the total number of predictions.

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

The AI4Bharat/IndicBERT model that was used for the vernacular languages produced the following results.

Label	precision	recall	f1-score
0	0.64	0.62	0.63
1	0.88	0.89	0.89

Key: 0: Homophobic or Transphobic ; 1: Non-Anti LGBTQ+. Performance metrics obtained for Tamil dataset using IndicBERT

The model produced the above results for the Tamil data set, with an accuracy of 0.83. The LaBSE model was used for the English dataset. The following results were obtained and an accuracy of 0.94 was procured.

Label	precision	recall	f1-score
0	0.57	0.16	0.25
1	0.95	0.99	0.97

[Key: 0: Homophobic or Transphobic ; 1: Non-Anti LGBTQ+] Performance metrics obtained for English dataset using LaBSE[10] model[1]

6. Error Analysis

Both the AI4Bharat/IndicBERT and LaBSE fail to attain perfect scores in identifying homophobic or transphobic comments (0 label). These scores were improved by making changes to the data preprocessing. Initially, when the emojis were completely removed from the text, the obtained scores were unsatisfactory. Thus, emoji replacement was done instead, as emoticons provide meaning and emotion to each sentence.

The IndicBERT model had difficulty always predicting the correct label. This Malayalam sentence, for instance, was incorrectly identified as a non-anti-LGBTQ+ statement. A possible explanation for this might be that the IndicBERT model is trained on native Indian language scripts and not their transliterated counterparts such as Malayalam sentences transliterated into English[11]. Hence, the model may be finding it difficult to identify the transliterated sentences in the corpus properly. One solution could be to try out models such as MuRIL which is trained on a corpus of traditional Indian scripts and their transliterated versions as well. This could be done for all the vernacular language datasets.

eg. Mal-06 - Mallu boy Boy avark purushan aayi purusha lingathodu koodi janichavark sthree hormon kooduthal aanenkilum oru poorna sthree aayi jeevikan orikkalum kazhiyilla athukond avante sareera ghadana enthano athinu anusrich manassine pakappeduthi jeevikuka allathe ith randum ketta oru vibhakamayi enthinu jeevikanam daivam 2 vibhakathine mathrame create cheythitullu aanineyum pennineyum

7. Conclusion

The paper discusses our approach to trying and identifying comments as homophobic, transphobic, or non anti LGBT. Our implementation included a mixture of IndicBert for the vernacular languages and LaBSE for English. The submitted IndicBert model got an F1 score of 0.228 for Tamil, 0.542 for Malayalam, and 0.393 for Tamil-English. The LaBSE submission received an F1 score of 0.4625.

On the whole, LaBSE seems to have performed better when compared to the IndicBERT model on the respective datasets. In the future, the performance could be improved by performing additional pre-processing steps on the data provided and perhaps, even making use of external datasets to improve the aforementioned accuracy of the transfer model implementations. As a future implementation, an ensemble approach could also be adopted to achieve a more accurate transfer model.

References

[1] 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.

[2] B. R. Chakravarthi, R. Priyadharshini, T. Durairaj, J. P. McCrae, P. Buitelaar, P. Kumaresan, R. Ponnusamy, 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, 2022, pp. 369–377.

[3] 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).

[4] J. B. PhD, J. S. PhD, S. Catalan, F. Gómez, J. Longueira, Discrimination and victimization: Parade for lesbian, gay, bisexual, and transgender (lgbt) pride, in chile, Journal of Homosexuality 57 (2010) 760–775. URL: https://doi.org/10.1080/00918369.2010.485880. doi:10.1080/00918369.2010.485880. arXiv:https://doi.org/10.1080/00918369.2010.485880, pMID: 20582801.

[5] C. G. Escobar-Viera, D. L. Whitfield, C. B. Wessel, A. Shensa, J. E. Sidani, A. L. Brown, C. J. Chandler, B. L. Hoffman, M. P. Marshal, B. A. Primack, For better or for worse? a systematic review of the evidence on social media use and depression among lesbian, gay, and bisexual minorities, JMIR mental health 5 (2018) e10496.

[6] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic bert sentence embedding, arXiv preprint arXiv:2007.01852 (2020).

[7] D. Kakwani, A. Kunchukuttan, S. Golla, G. N.C., A. Bhattacharyya, M. M. Khapra, P. Kumar, IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages, in: Findings of EMNLP, 2020.

[8] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A lite bert for self-supervised learning of language representations, arXiv preprint arXiv:1909.11942 (2019).

[9] B. Bharathi, J. Varsha, Ssncse_nlp@ tamilnlp-acl2022: Transformer based approach for emotion analysis in tamil language, in: Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, 2022, pp. 125–131.

[10] K. Swaminathan, K. Divyasri, G. Gayathri, T. Durairaj, B. Bharathi, Pandas@ abusive comment detection in tamil code-mixed data using custom embeddings with labse, in: Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, 2022, pp. 112–119.

[11] S. Bhawal, P. Roy, A. Kumar, Hate speech and offensive language identification on multilingual code mixed text using bert, in: Working Notes of FIRE 2021-Forum for Information Retrieval Evaluation (Online). CEUR, 2021.