Detection of Homophobia & Transphobia in Malayalam and Tamil: Exploring Deep Learning Methods

Deepawali Sharma[1](✉), Vedika Gupta[2], and Vivek Kumar Singh[1]

[1] Department of Computer Science, Banaras Hindu University, Varanasi, India
deepawali21@bhu.ac.in
[2] Jindal Global Business School, O.P. Jindal Global University, Sonipat, Haryana, India

Abstract. The increase in abusive content on online social media platforms is impacting the social life of online users. Use of offensive and hate speech has been making social media toxic. Homophobia and transphobia constitute offensive comments against LGBT + community. It becomes imperative to detect and handle these comments, to timely flag or issue a warning to users indulging in such behaviour. However, automated detection of such content is a challenging task, more so in Dravidian languages which are identified as low resource languages. Motivated by this, the paper attempts to explore applicability of different deep learning models for classification of the social media comments in Malayalam and Tamil languages as homophobic, transphobic and non-anti-LGBT + content. The popularly used deep learning models-Convolutional Neural Network (CNN), Long Short Term Memory (LSTM) using GloVe embedding and transformer-based learning models (Multilingual BERT and IndicBERT) are applied to the classification problem. Results obtained show that IndicBERT outperforms the other implemented models, with obtained weighted average F1-score of 0.86 and 0.77 for Malayalam and Tamil, respectively. Therefore, the present work confirms higher performance of IndicBERT on the given task on selected Dravidian languages.

Keywords: Deep learning · Homophobia · Malayalam · Tamil · Transphobia

1 Introduction

People share and post their views/opinions/thoughts about various topics on various social media platforms. The freedom to create such content also results in a large amount of displeasing data on racism, homophobia, transphobia, targeting particular communities and organizations, posted on the Web. Unfortunately, LGBT + people are being abused online by other communities resulting into problems of mental disorders (anxiety disorders, behavioural and emotional disorders, depression) in them. Sometimes, campaigns are initiated against these vulnerable communities when they come online for support and these campaigns are often converted into bullying [8]. Automatic detection of such content is a challenging task, more so if the text is in a low resource language.

I. Woungang et al. (Eds.): ANTIC 2022, CCIS 1798, pp. 217–226, 2023.
https://doi.org/10.1007/978-3-031-28183-9_15

Many studies have been done on the detection of hate speech [9–11], offensive and abusive content [12–14], cyberbullying [15–17] and many more on content in some selected languages. However, there is very little effort towards detecting racism, homophobia, transphobia in online texts from low resource Indian languages.

This paper presents an attempt towards this direction. Different deep learning-based models (CNN, LSTM, mBERT and IndicBERT) are implemented on data from two languages, namely Tamil and Malayalam. Both languages belong to the family of Dravidian languages and are low-resourced languages. The dataset provided by Dravidian-LangTech [8] is used for the purpose of the experiment. The comments are classified as Homophobic, Transphobic and Non-anti-LGBT + content. Those comments that contain pejorative labels or denigrative phrases against the gay, lesbian or any other community are classified as Homophobic [8]. People who can be homosexual, or can be transphobic without being homophobic, are identified as transphobic persons. Therefore, the comments that advocate or proposes anything unfavourable for transgender people are classified as transphobic [8]. Those comments that do not include homophobic, transphobic and have positive influence on readers are classified as non-anti-LGBT + content.

Different deep learning-based models are implemented to detect homophobia and transphobia in comments. More specifically, the paper performs the following tasks:

- Implementation of deep learning models (CNN, LSTM) with GloVe embedding for data in both languages (Malayalam and Tamil).
- Implementation of transformer-based models (mBERT, IndicBERT) for data in both languages (Malayalam and Tamil)
- Performance evaluation of the different models implemented on a standard dataset using standard metric.

The rest of the paper is organized as follows: Sect. 2 discusses the related work for the detection of homophobia and transphobia. Section 3 describes the dataset and its dimensions. Section 4 presents the experimental set up. Results are presented in Sect. 5. Section 6 concludes the paper with a brief summary of the results and major conclusions drawn.

2 Related Work

There is very limited previous work on detecting homophobia and transphobia in social media comments, particularly on low resource Indian languages. In one study, the ensemble transformer-based model was implemented to classify homophobia and transphobia on social media comments in Tamil and Tamil-English datasets [1]. Similarly, to classify the comments as homophobia and transphobia in English, different transformer-based models (BERT, RoBERTa and HateBERT) were implemented [2]. For detecting homophobia and transphobia, different monolingual and multilingual transformer-based models experimented on English, Tamil and Tamil-English datasets [3]. Another study experimented with transformer-based model (RoBERT-base) for English, Tamil and English-Tamil code-mixed datasets [6]. A study implemented neural network using sentence embedding and ensemble model in Tamil, English and Tamil-English [4]. Some of the studies implemented different classification algorithms using TF-IDF to classify the YouTube comments as homophobic and transphobic in English, Tamil and Tamil-English [5]. Similarly, the combination of word embeddings and Support Vector Machine (SVM) are implemented along with BERT-based models in English, Tamil and English-Tamil [7]. There is, however, no previous work on detecting homophobia and transphobia on social media comments in Malayalam language. Further, the previous studies have not explored the newer methods like IndicBERT. Therefore, the present study explores the suitability of various deep learning-based models, including the recently proposed IndicBERT model, for the task of detecting homophobia and transphobia on social media comments in Malayalam and Tamil languages.

3 Dataset Description

The dataset used in the study was launched by DravidianLangTech and is available online at https://codalab.lisn.upsaclay.fr/competitions/. The dataset contains the comments in two Dravidian languages namely Malayalam and Tamil. There are two attributes in the dataset: comment and category. The comment attribute consists of the sentiment/opinion of the people in form of text and the category tells that which of the three classes the comment belongs to: Homophobic, Transphobic, and Non-anti-LGBT + content. The Malayalam dataset contains 3,114 comments and the Tamil dataset contains 2,662 comments. Table 1 shows the category-wise distribution of comments in the Malayalam dataset. Table 2 shows the distribution of comments in each category of the Tamil dataset.

Table 1. Category-wise distribution of comments of Malayalam Dataset

Category	No. of comments
Homophobic	2434
Transphobic	491
Non-anti-LGBT + content	189

Table 2. Category-wise distribution of comments in Tamil Dataset

Category	No. of comments
Homophobic	2022
Transphobic	485
Non-anti-LGBT + content	155

3.1 Data Pre-processing

Since data pre-processing is a very important step in NLP applications and model building, we performed the following steps to pre-process the original text:

- Removing punctuations
- Removing Numbers
- URLs given in comments are removed.
- Repeated characters that do not add any information are removed.
- Extra blank spaces, that are created during pre-processing or if some comments already have that, are removed.

4 Experimental Setup

4.1 CNN

CNN was initially developed for image processing but in the past years, CNN showed good results in NLP tasks as well. When CNN is applied to text rather than the image, then 1-dimensional convolutional neural networks are used. To work with $1 - \text{D}$ CNN, a word embedding layer is required. GLoVe word embedding is used in this paper. For preparing the word embedding for the model, firstly all punctuations, links, etc. are removed, the text is tokenized and padding the text using the pad_sequences() to make equal length sequences. Figure 1 shows the implementation of CNN. From the word embedding, the words are represented as vectors. The embedding matrices are passed to the embedding layer. After that, the convolutional layer is added using "relu" as an activation function. Now, a pooling layer is added to reduce the complexity of the dimensions, and to be more specific MaxPooling is used. At last, the fully connected layer is added which is also called the dense layer in which the "Softmax" is used as an activation function since the dataset is multi-class classification. So, it classifies the comments into three categories: Homophobic, Transphobic and Non-anti-LGBT + content. The model is trained on 100 epochs for both languages (Malayalam and Tamil) and "Adam" is used as an optimizer. The loss is" Categorical Cross-entropy" to classify the tweets.

4.2 LSTM

LSTM is used for text classification because LSTM memorizes the information effectively and removes unwanted information. The LSTM model is able to find out the meaning of the input string and classify the text to the most precise output class if the appropriate layers of embedding are used. Before adding the embedding layer, firstly the data is pre-processed and the text is tokenized into tokens. After that, the sequence is padded by adding zero at the end of the sequence to make the sequence of the same length. Now, the word embedding (GLoVe) is used to represent the words into vectors. The padded sequence are passed to embedding layer as an input. The LSTM layer has been added after embedding layer using ReLu as an activation function. The dense layer is added at the end with activation function as softmax to classify the comments into more than two classes. Figure 2 shows the block diagram for LSTM to classify the comments. The model is trained on 100 epochs using the categorical cross-entropy as a loss function and "Adam" as an optimizer for both languages.

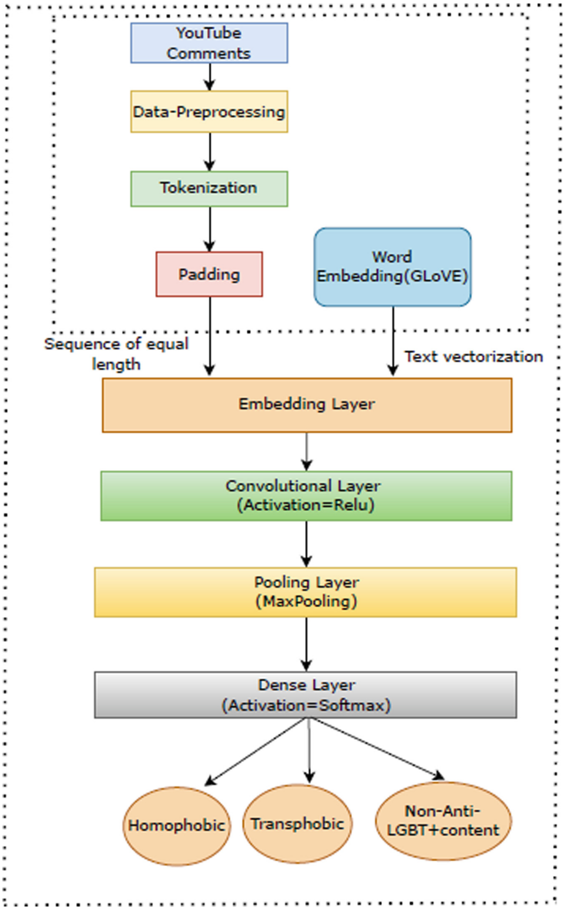

Fig. 1. Implementation of CNN to detect homophobia and transphobia

4.3 mBERT

BERT was developed in 2018 by Jacob Devlin and his colleagues from Google. BERT was trained on large-corpus of English data whereas mBERT was trained on 104 different languages including English [19]. It is not feasible to have separate BERT model for each language and therefore mBERT was introduced as a single BERT for 104 languages. In this paper, BERT base is used that has 12 layers of transformer encoder. Firstly, the comments are converted into tokens with the help of BERT tokenizer. Figure 3 shows the framework of BERT to classify the comments. Two special tokens [CLS] and [SEP] are added at the starting and end of the sequence, respectively. Vocabulary IDs for each of the tokens are token embeddings. Sentence embedding is used to distinguish between the sentences; and positional embedding is used to show the position of each word in the sequence. The model is build using the pre-trained BERT-base multilingual-cased

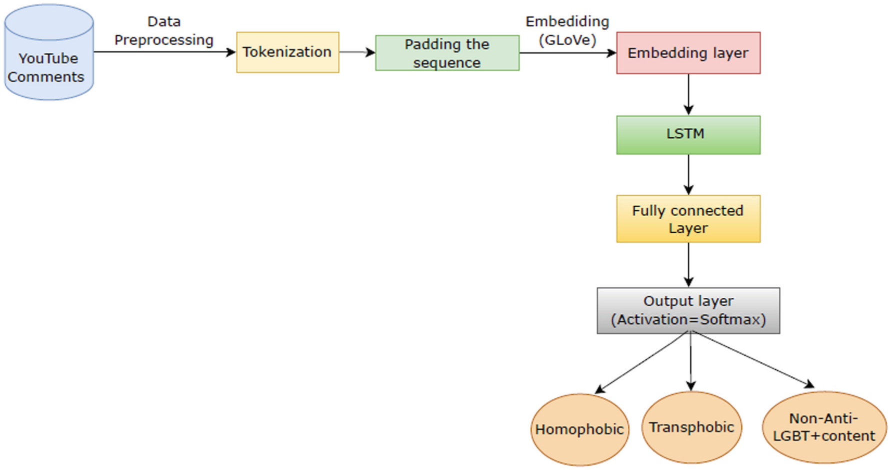

Fig. 2. Block diagram to show the implementation of LSTM model

model. The batch size is 32, learning is 3e-5 and the epochs = 5 is used to train the model on the dataset. At last the classifier classify the comments into homophobic, transphobic and non-anti-LGBT + content classes.

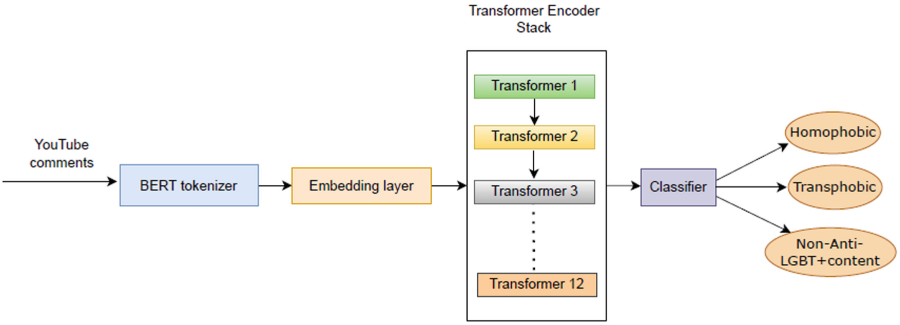

Fig. 3. Framework of transformer-based BERT model

4.4 IndicBERT

IndicBERT is a multilingual model trained on 12 Indian languages: Tamil, Malayalam, Bengali, Marathi, Punjabi, English, Hindi, Gujarati, Assamese, Oriya, Telugu and Kannada. As compared to other models like mBERT and XLM-R, it has a smaller number of parameters but even then gives the state-of-the-art (SOTA) performance for various tasks [18]. IndicBERT is based on the ALBERT model which is the derivative of the BERT. The text is tokenized using the Autotokenizer and the [CLS] token is added at the beginning of the sequence and [SEP] tokens are added at the end of the sequence similar to the BERT model. IndicBERT also has the ability to consider more than one sentence into a single sequence for input. As the dataset is in Indian languages: Malayalam and Tamil. For this dataset, the IndicBERT outperformed the other models and reported the weighted average F1 score for Malayalam and Tamil is 0.86 and 0.77, respectively. The parameters are fine-tuned: learning rate is set to be 3e-5, epochs = 5 and batch size is 32. At the end, the model classifies the comments into the three given categories: Homophobic, Transphobic and non-anti-LGBT + content.

5 Results

The CNN and transformer-based models are implemented to classify the comments from two languages into one of the three classes: 'Homophobic', 'Transphobic', and 'Non-anti-LGBT + content'. This section shows the performance of the implemented models. Table 3 shows the performance analysis of the CNN, LSTM and transfer learning models (mBERT and IndicBERT) for Malayalam. It can be observed that the weighted F1-score for CNN, LSTM, mBERT and IndicBERT are 0.80, 0.81, 0.83 and 0.86, respectively. For the Malayalam language, IndicBERT outperformed the other models. Table 4 shows the performance analysis of the implemented models for Tamil. The weighted average F1-score for deep learning models: CNN and LSTM are 0.71 and 0.72, respectively. The transformer-based model (IndicBERT) performs better than the other three implemented models (CNN, LSTM, mBERT), with a weighted F1 score of 0.77.

Table 3. Performance analysis for implemented models for Malayalam Dataset

Model	Homophobic			Transphobic			Non-anti-LGBT + content			Weighted average F1-score
	P	R	F1-score	P	R	F1-score	P	R	F1-score	
CNN(GloVe)	0.65	0.35	0.45	0.86	0.32	0.46	0.84	0.96	0.89	0.80
LSTM(GloVe)	0.69	0.45	0.54	0.60	0.32	0.41	0.86	0.95	0.90	0.81
mBERT	0.74	0.47	0.58	0.67	0.32	0.43	0.86	0.96	0.91	0.83
IndicBERT	0.79	0.49	0.59	0.70	0.39	0.50	0.88	0.97	0.91	**0.86**

* P = Precision, R = Recall

5.1 Comparison with Existing Studies

A comparative analysis of the deployed method is also done in order to understand the performance level of the method. Table 5 presents a comparative study of major previous studies on Tamil dataset with the performance obtained by the IndicBERT model. Studies experimented with transformer-based models. One study experimented with simple transformer model tamillion to classify the comments and reported weighted average F1-score is 0.75 [7]. Similarly, another study used RoBERTa-Base approach for the detection of comments into 3 classes: Homphobic, Transphobic and Non-anti-LGBT + content and obtained macro F1-score is 0.29 [6]. Similarly, the transformer- based

Table 4. Performance analysis for implemented models for Tamil Dataset

Model	Homophobic			Transphobic			Non-anti-LGBT + content			Weighted Average F1-score
	P	R	F1-score	P	R	F1-score	P	R	F1-score	
CNN(GloVe)	0.57	0.08	0.14	1.00	0.27	0.42	0.78	0.99	0.88	0.71
LSTM(GloVe)	0.67	0.12	0.21	0.80	0.27	0.40	0.79	0.98	0.87	0.72
mBERT	0.70	0.15	0.24	0.79	0.30	0.43	0.81	0.96	0.87	0.74
IndicBERT	0.72	0.18	0.28	0.76	0.34	0.46	0.84	0.94	0.88	**0.77**

* P = Precision, R = Recall

model (mBERT) experimented for the detection of homophobia and transphobia and reported F1-score is 0.64 [3]. There is no existing work on detecting homophobia and transphobia on social media comments in Malayalam language. The proposed approach (IndicBERT) is found to be outperforming other models and obtained weighted average F1-score for Malayalam and Tamil are 0.86 and 0.77 respectively. The obtained results of the IndicBERT model are better for both languages (Tamil and Malyalam) to classify the comments in given categories: homophobic, transphobic and non-anti-LGBT + content.

Table 5. Results of some of the previous studies

S. No.	Authors	Approach		F1-score	
		Tamil	Malayalam	Tamil	Malayalam
1	Swaminathan, K., Bharathi, B., Gayathri, G. L., & Sampath, H. (2022, May) [7]	Transformer monsoon-nlp/tamillion	N/A	0.75	N/A
2	Maimaitituoheti, A. (2022, May) [6]	RoBERTa-Base	N/A	0.29	N/A
3	Bhandari, V., & Goyal, P. (2022) [3]	mBERT	N/A	0.64	N/A
4	Proposed Approach (in this paper)	IndicBERT (Proposed Approach)	IndicBERT (Proposed Approach)	**0.77**	**0.86**

6 Conclusion

The paper presents the performance analysis of the different deep learning models (CNN, LSTM, mBERT and IndicBERT) on the task of homophobia and transphobia detection of the Malayalam and Tamil texts. The goal was to classify each comment into given three categories: Homophobic, Transphobic and Non-anti-LGBT + content. Experimental results show that the IndicBERT model outperformed the other models for both languages: Malayalam and Tamil, with reported weighted average F1-score of 0.86 and 0.77, respectively. While the other models also show reasonable performance, but the IndicBERT model is the best among all in terms of measured performance values. Thus, the IndicBERT model emerges as the superior of the various proposed methods for the tasks of Homophobia and Transphobia detection in the selected Dravidian languages. The present work can also be extended to data from other Indian languages. Further, ensemble-based models can also be explored for the given task.

Acknowledgement. This work is supported by an extramural research grant by HPE Aruba Centre for Research in Information Systems at BHU (No. M-22-69 of BHU).

References

1. Upadhyay, I.S., Srivatsa, K.A., Mamidi, R.: Sammaan@ lt-edi-acl2022: ensembled transformers against Homophobia and Transphobia. In: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pp. 270–275 (2022)
2. Nozza, D.: 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, pp. 258–264 (2022)
3. Bhandari, V., Goyal, P.: bitsa_nlp@ lt-edi-acl2022: leveraging pretrained language models for detecting homophobia and transphobia in Social Media Comments. arXiv preprint arXiv:2203.14267. (2022)
4. García-Díaz, J., Caparrós-Laiz, C., Valencia-García, R.: 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, pp. 140–144 (2022)
5. Ashraf, N., Taha, M., Abd Elfattah, A., Nayel, H.: 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, pp. 287–290 (2022)
6. Maimaitituoheti, A.: 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, pp. 155–160 (2022)
7. Swaminathan, K., Bharathi, B., Gayathri, G.L., Sampath, H.: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, pp. 239–244 (2022)
8. Chakravarthi, B.R., et al. Dataset for identification of homophobia and transophobia in multilingual YouTube comments. arXiv preprint arXiv:2109.00227
9. Khan, S., et al.: HCovBi-caps: hate speech detection using convolutional and Bi-directional gated recurrent unit with Capsule network. IEEE Access **10**, 7881–7894 (2022)
10. Khan, S., et al.: BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection. J. King Saud Univ.-Comput. Inform. Sci. **34**(7), 4335–4344 (2022)
11. Sap, M., Card, D., Gabriel, S., Choi, Y., Smith, N.A.: The risk of racial bias in hate speech detection. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp. 1668–1678 (2019)

12. Koufakou, A., Pamungkas, E. W., Basile, V., Patti, V.: HurtBERT: incorporating lexical features with BERT for the detection of abusive language. In: Fourth Workshop on Online Abuse and Harms, pp. 34–43. Association for Computational Linguistics (2020)
13. Susanty, M., Rahman, A.F., Normansyah, M.D., Irawan, A.: Offensive language detection using artificial neural network. In: 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), pp. 350–353. IEEE (2019)
14. Wiedemann, G., Ruppert, E., Jindal, R., Biemann, C.: Transfer learning from lda to bilstm-cnn for offensive language detection in twitter. arXiv preprint arXiv:1811.02906 (2018)
15. Cheng, L., Li, J., Silva, Y. N., Hall, D. L., Liu, H.: Xbully: Cyberbullying detection within a multi-modal context. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, pp. 339–347 (2019)
16. Balakrishnan, V., Khan, S., Fernandez, T., Arabnia, H.R.: Cyberbullying detection on twitter using Big Five and Dark Triad features. Personal. Individ. Differ. **141**, 252–257 (2019)
17. Hani, J., Mohamed, N., Ahmed, M., Emad, Z., Amer, E., Ammar, M.: Social media cyberbullying detection using machine learning. Int. J. Adv. Comput. Sci. Appl. **10**(5) (2019). https://doi.org/10.14569/IJACSA.2019.0100587
18. Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N. C., Bhattacharyya, A., Khapra, M. M., Kumar, P.: IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 4948–4961 (2020)
19. Devlin, J., Chang, M. W., Lee, K.,Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)