A Study on Sentimental Analysis, Homophobia-Transphobia Detection for Dravidian Languages

Manoj J Balaji[1], Chinmaya HS[2]

Abstract

With internet becoming highly accessible to mass population, there has been a tremendous increase in usage of social media, with the usage being spread across the Indian peninsula. Although this is advantageous, there's also increase in anti-social activities in the social media space. There has been an increase in hate speech especially the ones that lie in the spectrum of homophobia and trans-phobia. With a growing concern for preventing such posts on the social media, there are multiple efforts happening across the world. To solve this issue, we study two methods, fastText+LightGBM based classification for Sentimental analysis and MP-Net is used for homophobia-trans-phobia detection. For this study, we are using the dataset provided by the shared task on Sentiment Analysis and Homophobia detection of YouTube comments in Code-Mixed Dravidian Languages. The proposed methodology for sentimental analysis has macro f1 scores of 0.19, 0.3, 0.2 for Tamil, Kannada and Malayalam respectively and for homophobia-transphobia detection, the macro f1 scores are 0.234, 0.493, 0.942, 0.316 for Tamil, English, Malayalam and Tamil-English respectively. The proposed solution outshines baselines for homophobia-transphobia detection.

Keywords

Homophobia Detection, Transphobia Detection, Sentiment Analysis, Social Media, Dravidian Language, MP-Net, Classification, Transformers, LightGBM,

1. Introduction

In the recent advancement of technology and social media hatred towards LGBTQ+ community is also growing. Homophobia/transphobia refers to the actions resulting in threat, dread, dislike, discomfort or mistrust of lesbian, gay, transgender or bisexual person [1]. Social media, as it provides medium for communication, allowing the users to express their views, ideas and feelings on anything at any time. The power of sharing resources, materials to support their views are also enabled using social media platforms [2] [3]. The abundance of data available online can enable researchers to use natural language processing to interpret, quantify, and monitor the user behavior, propagation of information across different communities and events influence by these online information [4].

Internet is home to a wide verity of racist, sexist, homophobic, trans phobic and all sorts of unpleasant content.The increase in the quantity of such contents have appeared as a problem for online communities [5]. The wide verity of data available on social media platforms such

Forum for Information Retrieval Evaluation, December 9-13, 2022, India

manojbalaji1@gmail.com (M. J. Balaji); chinmayasbhat4@gmail.com (C. HS)

as YouTube, Facebook and others are ever changing and are influencing the way people think, talk and connect with each other.Social media platforms also provide a great avenue to venture into the darker side of internet, like, share and support the violent, sexist, homophobic content creation and sharing [6].

With increasing content available on the internet, the computer scientists, linguists and researchers have an opportunity to build and use automated solutions that can mitigate or ban anti LGBTQ+ harmful content, and try to make internet a place of equality, diversity and inclusion. While much work has been put into the domain of aggression identification [7], misogyny [8] [5], and racism [9], homophobic or transphobic verbal abuse, on the other hand, was given as far less important than racist or other prohibited issues

Recent advancements in the attention mechanism used in transformers, which are becoming very popular in low resource Dravidian languages like, Tamil, Malayalam, Kannada among others. Lack of language carpus available to train makes it difficult to train the models without using embedding where transformers are acting as a solution. Bert [10] and XLnet [11], which are the two highly popular models used for the text classification and which are pose to have drawbacks which are overcome with MPNet.

2. Related Work

Wast availability of data on the internet attracted many researchers and computer scientist to develop and research possible solutions to tackle the hatred towards LGBTQ+ communities. One of the early studies towards identifying offensive comment identification in dravidian languages (Tamil) [12] [13] followed by DravidianLangtech [14] shed light towards possibilities in bringing equality and diversity for LGBTQ+ people who are also ill-treated in these part. Dataset for HASOCDravidianCodeMix which consisted of 4000 comments which were collected from twitter and other social media platforms. Similar work DravidianLangTech comprised of 30 thousand YouTube comments, which were annotated by multiple volunteers. Both these datasets are code mixed Datasets. Based on these two datasets k [14]. Inspired by these works, our previous work on dravidian code mix dataset (troll-meta) [15], created a hybrid deep learning model which performed classification of given images to one of the 2 classes. The work focused on classification of data into offensive speech and neutral ones.

Works by Ljubešić et al. [16] constructed lexicons of several languages including Croatian, Dutch and Slovene. And using these lexicons to identify texts containing socially unacceptable words towards topics of migrants and LGBTQ+. Even though this is a great work, but it fails to meet the end goals as it was in the early stages of research, it lacks confidence in classification task.

DravidianCodeMix, a recent work proposes a multilingual model which tries to establish a baseline model to conduct further research [1]. The corpera included comments collected from Youtube, belonging to 4 major dravidian languages, Kannada, Tamil and Malayalam. The data set has Kannada-English, Tamil-English and Malayalam-English datasets.Which are annotated by human volunteers.

Our work focus on sentiment analysis and classification of homophobic and transphobic comments that are collected by YouTube.

3. Dataset

The dataset is provided as part of shared task on "Sentiment Analysis and Homophobia detection of YouTube comments in Code-Mixed Dravidian Languages" [17]. The data set consists of annotated data for sentiment analysis and offensive language identification for a total of more than 60 thousand individual comments on YouTube videos.

The dataset count for individual languages and tasks are tabulated in Tables 1, 2, 3 and 4.

	Tamil	Kannada	Malayalam
Positive	20069	2823	6421
Negative	4271	1188	2105
Mixed Feeling	4020	574	926

Table 1
Sentiment Analysis - Train Data

For sentiment analysis the dataset belonging Tamil language contained 20069 positive, 4271 negative and 4020, Kannada language contained 2823 positive, 1188 Negative and 574 Mixed feeling, and Malayalam language data consisted of 6421 positive, 2105 negative and 926 mixed feeling comments. All these are labeled by the volunteers as mentioned in [18]

	Tamil	Kannada	Malayalam
Positive	2257	321	706
Negative	480	139	237
Mixed Feeling	438	52	102

Table 2
Sentiment Analysis - Dev Data

The development dataset, which contained 2257 positive, 480 negative and 438 Mixed feeling data for Tamil language, 321 positive, 139 negative and 52 mixed feeling data for Kannada Language and for Malayalam there are 706 positive, 237 negative and 102 mixed feeling.

	Tamil	English	Malayalam	Tamil-English
Non-anti-LGBT+ content	2022	3001	2434	3438
Homophobic	485	157	491	311
Transphobic	155	6	189	112

Table 3
Homophobia/Transphobia Analysis - Train Data

Homophobia/Transphobia detection training dataset contained 2022 Non-anti-LGBTQ+ content, 485 Homophobic and 155 Transphobic data for Tamil, 3002 Non-anti-LGBTQ+ content, 157 Homophobic and 6 Transphobic data for English, for language Malayalam, 2434 Non-anti-LGBTQ+content, 491 homophobic and 189 transphobic comments and finally for Tamil-English 3438 non-anti-LGBTQ+ content, 311 homophobic and 112 Transphobic comments.

	Tamil	English	Malayalam	Tamil-English
Non-anti-LGBT+ content	526	732	692	862
Homophobic	103	58	133	66
Transphobic	37	2	41	38

Table 4
Homophobia/Transphobia Analysis - Dev Data

The dev dataset for Homophobia/Transphobia detection contained 526 Non-anti-LGBTQ+ content, 103 homophobic and 37 transphobic comments for Tamil, 532 non-anti-LGBTQ+ content, 58 homophobic and 2 transphobic comments for English, for Malayalam 692,133 and 41 comments for Non-anti-LGBTQ+ content, homophobic and transphobic labels respectively. Tamil-English dev data consisted of 862 non-anti-LGBTQ+ content, 66 homophobic and 38 transphobic comments.

4. Approach

The approach to the solution started with cleaning the data, making it free from special characters, converting the Kannada-English, Tamil-English and Malayalam-English to lowercase.

Emoji's as the name suggests, which is used to express emotions in the form of graphics, images, pictogram or ideogram embedded with text. We consider these as one of the major driver in finding emotions such as sarcasm, sadness, happiness etc. They play major role in finding or classifying emotions and analysing the sentiments in the given comments. We used the Python library (https://pypi.org/project/emoji/) to convert the emoji's to text.

For Example:
Original: Idha pathutu road la students kathi vaichi kitu sanda poduvanunga 🤦
Translated: Idha pathutu road la students kathi vaichi kitu sanda poduvanunga Face Palm
The study involves 2 tasks which are Sentimental Analysis and Homophobia-Transphobia Detection which will be referred to as Task A and Task B henceforth. Task A involved 3 different languages which are Tamil, Kannada, and Malayalam whereas for Task B, 4 languages i.e. Tamil, English, Malayalam, and Tamil-English(combination of both, often called as Tanglish in colloquial language) were in consideration.

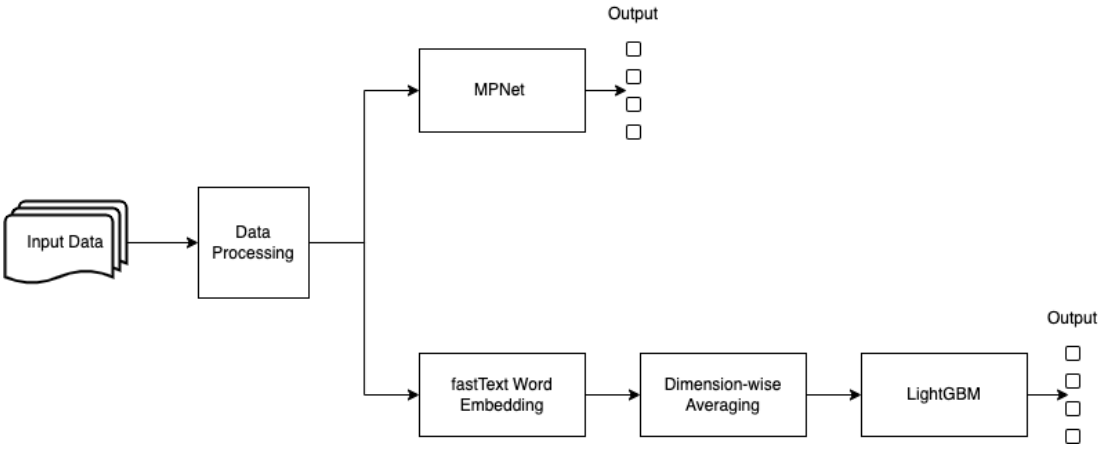

Figure 1: Architecture of the proposed system

With the help of open source libraries for MPNet, fastText and transformers with the fine tuning of hyper-parameters like learning rate, weight decay, batch size along with others.

For both the tasks, two methods were experimented. The first method involved building a text classifier using MP-Net [19] and the second method involved word generating embedding using fastText [20], followed by dimension-wise averaging, finally classifying using LightGBM [21] to obtain the required results.

For LightGBM, 15 num_leaves, min_child_weight of 1e-1, subsample of 0.8 and random state of 42 give the better results for the task of Homophobia/Transphobia detection.

Similar to this, MPNet trained with a learning rate of 2e-5, weight decay of 0.01 and batch size of 8 for the task of sentiment analysis.

For Sentimental Analysis, MP-Net method was used, and for Homophobia-Transphobia Detection, for Tamil and Malayalam, fastText+LightGBM method was used whereas for English and Tamil-English, MP-Net was used.

5. Results

The research activity performed for sentiment analysis as well as classifying the comments into homophobic/transphobic comments or not. The model involved two different machine learning model for the classification problem, part-A is a decision tree based LightGBM [21], where as the part-B was a hybrid model of masked language modeling and permuted language modeling [19].

To analyze the results of the model, a confusion matrix was constructed, and the weighted f1 is calculated.

The results for the conducted research activity are tabulated in the table 5 and 6.

	precision	recall	F1-score	Rank
Tamil	0.190	0.220	0.190	9
Kannada	0.290	0.330	0.300	12
Malayalam	0.160	0.270	0.200	2

Table 5
Results for Sentiment Analysis

	Macro F1-score	Rank
Tamil	0.234	5
English	0.493	1
Malayalam	0.942	3
Tamil-English	0.316	8

Table 6
Results for Homophobia/Transphobia Detection

6. Error Analysis

The research activity carried out shed the light towards the setbacks faced during training and evaluation steps. One of the important reason for the performance of the models is lack of data pertaining to dravidian languages compared to other, as it can be seen in the result of classifying homophobia/transphobia. Wherein, the rank in English task is 1, which is because the MPNet is trained on larger English language data corpus. That is the reason why we explored fastText as an alternate option. The other models in the same task were trained on fastText+LightGBM, even though fastText were also trained on Tamil/Malayalam language corpus, due to the differences in colloquial language versus formal language in which the models were trained on, the results were poor. Tamil-English performed poor in ranking compared to others, which is likely as pre-trained models seldomly comes across language code-switch, thus failing to provide better representation embedding.

Dataset size is another aspect that we analyze the results. The size of the dataset available for Tamil language is more compared the others, due to which performance is better, which can be clearly seen in the results. With the minimum precision, Malayalam language had least quantity of data next to Kannada. Even with embeddings from transformers, the quantity of data were not enough for better generalization.

In terms of of improvement, pre-training or fine-tuning these aforementioned models on the available dataset, will significantly increase the quality of predictions. Also we will have to explore other methodologies to handle low-resource constraints and strive to achieve best results.

7. Conclusion

We experimented with both fastText along with LightGBM and MPNet which were able to provide some improvements over the baseline models. Even with considerable improvements the models experienced some shortcomings [22][18].

In the future works we are considering building custom transformers and enhanced architectures to gain a better results.

References

[1] 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, 2021. URL: https://arxiv.org/abs/2109.00227. doi:10.48550/ARXIV.2109.00227.

[2] J. Qian, A. Bethke, Y. Liu, E. Belding, W. Y. Wang, A benchmark dataset for learning to intervene in online hate speech, 2019. URL: https://arxiv.org/abs/1909.04251. doi:10.48550/ARXIV.1909.04251.

[3] G. Gkotsis, A. Oellrich, T. Hubbard, R. Dobson, M. Liakata, S. Velupillai, R. Dutta, The language of mental health problems in social media, in: Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology, Association for Computational Linguistics, San Diego, CA, USA, 2016, pp. 63–73. URL: https://aclanthology.org/W16-0307. doi:10.18653/v1/W16-0307.

[4] K. Yamada, R. Sasano, K. Takeda, Incorporating textual information on user behavior for personality prediction, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Association for Computational Linguistics, Florence, Italy, 2019, pp. 177–182. URL: https://aclanthology.org/P19-2024. doi:10.18653/v1/P19-2024.

[5] S. Bhattacharya, S. Singh, R. Kumar, A. Bansal, A. Bhagat, Y. Dawer, B. Lahiri, A. K. Ojha, Developing a multilingual annotated corpus of misogyny and aggression, 2020. URL: https://arxiv.org/abs/2003.07428. doi:10.48550/ARXIV.2003.07428.

[6] H. Mulki, B. Ghanem, Let-mi: An arabic levantine twitter dataset for misogynistic language, 2021. URL: https://arxiv.org/abs/2103.10195. doi:10.48550/ARXIV.2103.10195.

[7] J. Risch, R. Krestel, Aggression identification using deep learning and data augmentation, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 150–158. URL: https://aclanthology.org/W18-4418.

[8] E. Fersini, D. Nozza, G. Boifava, Profiling Italian misogynist: An empirical study, in: Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 9–13. URL: https://aclanthology.org/2020.restup-1.3.

[9] Z. Waseem, Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter, in: Proceedings of the First Workshop on NLP and Computational

Social Science, Association for Computational Linguistics, Austin, Texas, 2016, pp. 138–142. URL: https://aclanthology.org/W16-5618. doi:10.18653/v1/W16-5618.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL: https://arxiv.org/abs/1810.04805. doi:10.48550/ARXIV.1810.04805.

[11] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, 2019. URL: https://arxiv.org/abs/1906.08237. doi:10.48550/ARXIV.1906.08237.

[12] B. R. Chakravarthi, HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion, in: Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media, Association for Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 41–53. URL: https://aclanthology.org/2020.peoples-1.5.

[13] T. Mandl, S. Modha, A. Kumar M, B. R. Chakravarthi, Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german, in: Forum for Information Retrieval Evaluation, FIRE 2020, Association for Computing Machinery, New York, NY, USA, 2020, p. 29–32. URL: https://doi.org/10.1145/3441501.3441517. doi:10.1145/3441501.3441517.

[14] B. R. Chakravarthi, R. Priyadharshini, N. Jose, A. Kumar M, T. Mandl, P. K. Kumaresan, R. Ponnusamy, H. R L, J. P. McCrae, E. Sherly, Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, 2021, pp. 133–145. URL: https://aclanthology.org/2021.dravidianlangtech-1.17.

[15] M. B. J, C. Hs, TrollMeta@DravidianLangTech-EACL2021: Meme classification using deep learning, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, 2021, pp. 277–280. URL: https://aclanthology.org/2021.dravidianlangtech-1.39.

[16] N. Ljubešić, I. Markov, D. Fišer, W. Daelemans, The LiLaH emotion lexicon of Croatian, Dutch and Slovene, in: Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media, Association for Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 153–157. URL: https://aclanthology.org/2020.peoples-1.15.

[17] K. Shanmugavadivel, M. Subramanian, P. K. Kumaresan, B. R. Chakravarthi, B. B, S. Chinnaudayar Navaneethakrishnan, L. S.K, T. Mandl, R. Ponnusamy, V. Palanikumar, M. B. 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.

[18] 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).

[19] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, Mpnet: Masked and permuted pre-training for language understanding, 2020. URL: https://arxiv.org/abs/2004.09297. doi:10.48550/

ARXIV.2004.09297.

[20] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information (2016). URL: https://arxiv.org/abs/1607.04606. doi:10.48550/ARXIV.1607.04606.

[21] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

[22] B. R. Chakravarthi, R. Priyadharshini, T. Durairaj, J. 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, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 369–377. URL: https://aclanthology.org/2022.ltedi-1.57. doi:10.18653/v1/2022.ltedi-1.57.