DE-ABUSE@TamilNLP-ACL 2022: Transliteration as Data Augmentation for Abuse Detection in Tamil

Vasanth Palanikumar[1], **Sean Benhur**[2], **Adeep Hande** [3]
Bharathi Raja Chakravarthi[4]
[1]Chennai Institute of Technology [2]PSG College of Arts and Science
[3] Indian Institute of Information Technology Tiruchirappalli
[4]National University of Ireland Galway
vasanthpcse2019@citchennai.net, seanbenhur@gmail.com
adeeph18c@iiitt.ac.in, bharathi.raja@insight-centre.org

Abstract

With the rise of social media and internet, there is a necessity to provide an inclusive space and prevent the abusive topics against any gender, race or community. This paper describes the system submitted to the ACL-2022 shared task on fine-grained abuse detection in Tamil. In our approach we transliterated code-mixed dataset as an augmentation technique to increase the size of the data. Using this method we were able to rank 3rd on the task with a 0.290 macro average F1 score and a 0.590 weighted F1 score.

1 Introduction

Internet is a global computer network that provides a variety of information and facilitates communication between users from any part of the world. The world population is 7.9 billion as of January 2022 of which around 5.2 billion are live internet users[1]. In recent times, people have become more communicative and inclusive. People want to share their views on a common platform, where social media comes into the picture (Chakravarthi, 2020; Chakravarthi and Muralidaran, 2021; Puranik et al., 2021; Ghanghor et al., 2021). People can post their opinions which are productive and efficient for their society but at times people also post their opinions which could be abusive to others. There are many social media platforms like YouTube, Facebook, Instagram, Twitter and many more (Priyadharshini et al., 2021; Kumaresan et al., 2021) where the users are given the liberty to put forward their opinion. On an average as per statistics around 250 M tweets are posted, 2 million blogs are written on various websites and 80 B mails are sent per day. Social media platforms could be both a boon and bane.

Comments that humiliates or denigrates an individual or a group based on various characteristics such as colour, ethnicity, sexual orientation, nationality, race and religion are called abusive comments(Saumya et al., 2021). Abuse caused via social media can cause many negative impacts in users' lives. This will affect the mental state of the specific individual terribly causing depression and sleeplessness (Chakravarthi et al., 2021c; Sampath et al., 2022; Chakravarthi et al., 2022). Some of these comments also can create a controversy over the social media on a specific individual or a group of people. This shows the need for restricting these kind of abusive comments from being posted in the social media. Once abusive comments have been posted onto the social media it should be flagged and immediately removed.

This world is a diverse one which comprises of different kinds of people from different origin. But when it comes to the comments of people the language plays a very important role ("Bharathi et al., 2022). Though most of the people use English as their language to show their opinion some of them also use other languages instead of English. For example in a diverse nation like India where people are not restricted to communicate in English, people comment in different languages like Tamil, Telugu, Kannada, Malayalam, Hindi, Marathi and many others.

Tamil is one of the oldest and longest surviving language in this world(Chakravarthi et al., 2020). It is an old Dravidian language mostly spoken by people of South Indian origin with a history of over 3000 years[2] that has lot of dialects. Therefore it is very tough to classify posts which have abusive comments in Tamil language.

Lately, after the advent of machine learning, researches are carried over onto this area for classifying the abusive comments.

In our work we have used Transformer (Vaswani et al., 2017) models for the given task of classifying the abusive comments. The rest of the paper is

[1]https://www.internetlivestats.com/

[2]https://www.cal.org/heritage/pdfs/Heritage-Voice-Language-Tamil.pdf

Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pages 33 - 38

structured as follows section 2 describes the related works which are carried over in this field. The section 3 describes the methodology used in the system. The 2nd section describes the results we obtained in our research experiments.We discuss our results on Section 4. Finally in the 5th section we conclude this research paper followed by the references section.

2 Related Work

2.1 NLP on Tamil

NLP in Tamil have been recently carried out extensively through various shared tasks(Chakravarthi et al., 2021b,a) focusing on tasks such as offensive language detection, machine translation and sentiment analysis. Participants have used different methods including intelligent feature extraction (Dave et al., 2021) and ensembles of deep learning methods (Saha et al., 2021). Tamil is an agglutinative language, due to the ease of typing many users use Tamil in roman script in the social media and internet, this is known as code-switching (Jose et al., 2020), since it is also a morphologically rich language, developing NLP systems in Tamil is hard.

2.2 Abuse detection

Tasks such as abuse detection, offensive language detection and hate speech detection have been a focus of research for the past decade due to a surge in the internet and social media platform users. With the emergence of deep learning and transformers, current approaches for abusive language detection heavily relies on deep learning methods due to the rise of transformers and pretrained language models, since pretrained language models require less data.

3 Methodology

In this section, we describe the methodology based on which our system is designed, including the data preparation phase, modelling phase and model evaluation phase.

3.1 Data Preprocessing

In the shared task, two datasets (Priyadharshini et al., 2022) were provided where one comprises of Tamil sentences while the other comprising of code-mixed Tamil-English sentences. The Tamil dataset comprises of 2,240 sentences for training and 560 sentences for validation. In the code-mixed dataset there are 5,948 training sentences and 1,488 validation sentences.Table 1 shows the distribution of data among different classes before and after combining Tamil and Transliterated dataset.

We first removed punctuations present in both the dataset. The datasets comprises of some categories like Transphobic there were only very few sentences corresponding to it. To overcome this data shortage issue we performed transliteration on the code-mixed dataset and we converted the sentences in that dataset also to its corresponding Tamil sentences (Hande et al., 2021) by using ai4bharat-transliteration [3] Python package. Before combining the dataset, we removed all those sentences which fell under the category of not-Tamil and then combined the Tamil dataset with the transliterated dataset ending up with 8,186 sentences which is approximately 4 times the size of the previous dataset. By this the imbalance in the dataset was reduced and we overcame the data-shortage as well.

Figure 1 depicts the data preparation phase graphically.

3.1.1 Transliteration

Transliteration refers to the process of converting a word from one script to another wherein the semantic meaning of the sentence is not changed and the syntactical structure of the target language is strictly followed (Hande et al., 2021). By this we have increased our data size considerably. For this Transliteration we have used ai4bharat-transliteration Python package.

3.2 Modelling

In our experimentation, MURIL model outperformed all the other models which we experimented on. For evaluation we considered macro and weighted F1-score.

3.2.1 ML Models with N-gram TF-IDF Vectorization

For experimenting with ML models, we created a pipeline where first the text is vectorized by using CountVectorizer and is transformed by TfIdfTransformer. Once the transformation of the data is completed, it is trained on the following Machine Learning models: LightGBM, Catboost, RandomForest, Support Vector Machines classifer and Multinomial Naive Naive Bayes. Of the all models

[3] https://github.com/AI4Bharat/IndianNLP-Transliteration

Classes	Tamil Dataset	Transliterated dataset	Combined dataset
Counter-speech	149	348	497
Homophobia	35	172	207
Hope-Speech	86	213	299
Misandry	446	830	1276
Misogyny	125	211	336
None-of-the-above	1296	3715	5011
Transphobic	6	157	163
Xenophobia	95	297	392

Table 1: Distribution of Dataset

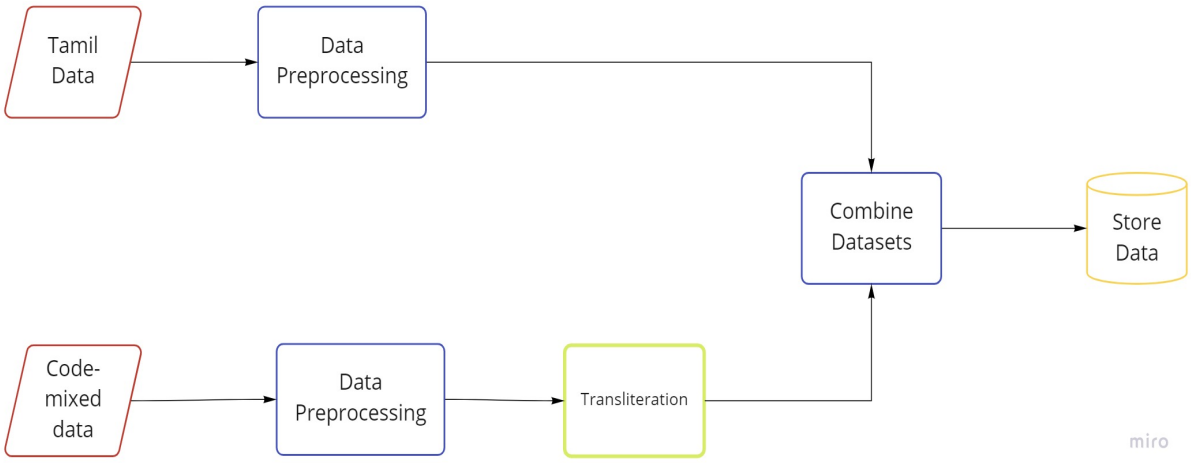

Figure 1: Data Preparation phase

experimented LightGBM (Ke et al., 2017) outperformed all the other algorithms by having 0.32 macro average f1-score and 0.65 weighted average f1-score followed by Catboost. Therfore we performed hyperparameter tuning on Optuna on LightGBM where we ended up having 0.36 macro average f1-score and 0.63 weighted average f1-score which was the highest metric of our experiments on traditional ML models.

3.2.2 MURIL

MURIL (Khanuja et al., 2021) is a pretrained bert model created by Google for tasks on Indian languages trained on 17 Indian languages. It was parallely trained on Translated Data and Transliterated Data. Based on the XTREME (Hu et al., 2020) benchmark, MURIL outperformed mBERT for all the languages in all standard downstream tasks. Hence, this model handles translated and transliterated data very well. We fine-tuned the MURIL model with the parameters listed in the Table 3. The metric we obtained from MURIL showed us that it outperformed all other ML models.

4 Results

MURIL and other Machine Learning models were trained on the training set and was validated on the dev set. For this competition, submission, macro f1-score was considered as the metric of evaluation by the organisers. By this MURIL trained on both Tamil and Transliterated dataset combined together had a very high macro f1-score of 0.49 and weighted f1-score of 0.76 on the validation dataset and a macro f1-score of 0.290 on test dataset and weighted f1-score of 0.590. With this result we secured the 3^{rd} rank in the task. The Table 2 shows the results of all the experimentations carried on during the modelling phase.

5 Conclusion

In this paper, we conclude that with a relatively smaller-size dataset, we can use Transliteration as an efficient data augmentation technique to increase the volume of data available which played a very important role for getting a better F1-score is evident from the results Table 2 shows that Transliteration of dataset works very well. We also con-

Model	Dataset	MP	MR	MF	WP	WR	WF
MURIL	Tamil	0.37	0.34	0.33	0.67	0.70	0.68
MURIL	Combined	0.52	0.44	0.46	0.72	0.72	0.71
LightGBM	Tamil	0.30	0.42	0.33	0.76	0.65	0.69
LightGBM	Combined	0.28	0.46	0.31	0.78	0.66	0.71
CatBoost	Tamil	0.28	0.52	0.33	0.82	0.66	0.72
CatBoost	Combined	0.26	0.43	0.29	0.82	0.66	0.72
Random Forest	Tamil	0.23	0.55	0.25	0.81	0.63	0.70
Random Forest	Combined	0.23	0.39	0.24	0.85	0.65	0.73
Support Vector Machine	Tamil	0.24	0.53	0.26	0.87	0.65	0.73
Support Vector Machine	Combined	0.26	0.45	0.28	0.88	0.66	0.75
Multinomial Naive Bayes	Tamil	0.16	0.16	0.14	0.94	0.64	0.74
Multinomial Naive Bayes	Combined	0.18	0.29	0.18	0.93	0.65	0.75

Table 2: Experimental Results on various models **MF** - macro F1-score; **WF** - weighted F1-score; **MP** - macro Precision; **WP** - weighted Precision; **MR** - macro Recall; **WR** - weighted Recall

Hyperparameters	Values
Learning Rate	2e-5
Batch Size	16
Epochs	3
Weight Decay	0.001
Dropout	0.3

Table 3: Hyperparameters used across experiments

clude that Transformer models outperform traditional Machine Learning and Deep Learning models for this task.

References

B "Bharathi, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, N Sripriya, Arunaggiri Pandian, and Swetha" Valli. 2022. Findings of the shared task on Speech Recognition for Vulnerable Individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi. 2020. 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*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Dhivya Chinnappa, Ruba Priyadharshini, Anand Kumar Madasamy, Sangeetha Sivanesan, Subalalitha Chinnaudayar Navaneethakrishnan, Sajeetha Thavareesan, Dhanalakshmi Vadivel, Rahul Ponnusamy, and Prasanna Kumar Kumaresan. 2021a. Developing successful shared tasks on offensive language identification for dravidian languages.

Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran. 2021. Findings of the shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 61–72, Kyiv. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

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, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John P. McCrae, and Elizabeth Sherly. 2021b. 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*, pages 133–145, Kyiv. 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. 2021c. Dataset for identification of homophobia and transophobia in multilingual youtube comments. *arXiv preprint arXiv:2109.00227*.

Bhargav Dave, Shripad Bhat, and Prasenjit Majumder. 2021. IRNLP_DAIICT@DravidianLangTech-EACL2021:offensive language identification in Dravidian languages using TF-IDF char n-grams and MuRIL. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 266–269, Kyiv. Association for Computational Linguistics.

Nikhil Ghanghor, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITK@LT-EDI-EACL2021: Hope speech detection for equality, diversity, and inclusion in Tamil , Malayalam and English. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 197–203, Kyiv. Association for Computational Linguistics.

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. *CoRR*, abs/2108.12177.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Navya Jose, Bharathi Raja Chakravarthi, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2020. A survey of current datasets for code-switching research. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 136–141.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *NIPS*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha P. Talukdar. 2021. Muril: Multilingual representations for indian languages. *CoRR*, abs/2103.10730.

Prasanna Kumar Kumaresan, Ratnasingam Sakuntharaj, Sajeetha Thavareesan, Subalalitha Navaneethakrishnan, Anand Kumar Madasamy, Bharathi Raja Chakravarthi, and John P McCrae. 2021. Findings of shared task on offensive language identification in tamil and malayalam. In *Forum for Information Retrieval Evaluation*, pages 16–18.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumar Kumaresan. 2022. Findings of the shared task on Abusive Comment Detection in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Sajeetha Thavareesan, Dhivya Chinnappa, Durairaj Thenmozhi, and Rahul Ponnusamy. 2021. Overview of the dravidiancodemix 2021 shared task on sentiment detection in tamil, malayalam, and kannada. In *Forum for Information Retrieval Evaluation*, pages 4–6.

Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITT@LT-EDI-EACL2021-hope speech detection: There is always hope in transformers. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 98–106, Kyiv. Association for Computational Linguistics.

Debjoy Saha, Naman Paharia, Debajit Chakraborty, Punyajoy Saha, and Animesh Mukherjee. 2021. Hatealert@DravidianLangTech-EACL2021: Ensembling strategies for transformer-based offensive language detection. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 270–276, Kyiv. Association for Computational Linguistics.

Anbukkarasi Sampath, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Ruba Priyadharshini, Subalalitha Chinnaudayar Navaneethakrishnan, Kogilavani Shanmugavadivel, Sajeetha Thavareesan, Sathiyaraj Thangasamy, Parameswari Krishnamurthy, Adeep Hande, Sean Benhur, and Santhiya Ponnusamy, Kishor Kumar Pandiyan. 2022. Findings of the shared task on Emotion Analysis in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Sunil Saumya, Abhinav Kumar, and Jyoti Prakash Singh. 2021. Offensive language identification in Dravidian code mixed social media text. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 36–45, Kyiv. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.