HARMONY@DravidianLangTech: Transformer-based Ensemble Learning for Abusive Comment Detection

Amrish Raaj P, Abirami S, Lysa Packiam R S, Deivamani M
Department of Information Science and Technology
College of Engineering Guindy
Anna University, Chennai, India
amrishraaj@gmail.com, abirami@auist.net, mailtolysa@gmail.com
deivamani@auist.net

Abstract

Millions of posts and comments are created every minute as a result of the widespread use of social media and easy access to the internet. It is essential to create an inclusive environment and forbid the use of abusive language against any individual or group of individuals.This paper describes the approach of team HARMONY for the "Abusive Comment Detection" shared task at the Third Workshop on Speech and Language Technologies for Dravidian Languages.A Transformer-based ensemble learning approach is proposed for detecting abusive comments in code-mixed (Tamil-English) language and Tamil language. The proposed architecture achieved rank 2 in Tamil text classification sub task and rank 3 in code mixed text classification sub task with macro-F1 score of 0.41 for Tamil and 0.50 for code-mixed data.

1 Introduction

Online Social Networks (OSNs) have become increasingly significant in recent years and are now considered to be go-to resource for news, knowledge, and entertainment (Halevy et al., 2022; Priyadharshini et al., 2021; Kumaresan et al., 2021) . However, despite the numerous benefits of using OSNs, mounting evidence suggests that an increasing number of illicit actors are taking advantage of these platforms to spread toxicity and harm other people. Numerous risks are brought on by these illicit people, including online abuse, vulgarity, harassment and bullying.

Abusive comments are those that mock or disparage a person or a group based on traits like colour, ethnicity, sexual orientation, nationality, race, or religion (Saumya et al., 2021). Social media abuse can have a negative impact on users' lives in a number of ways. This will have a terrible impact on that person's mental health, leading to depression and insomnia (Chakravarthi et al., 2021; Sean, 2022; Chakravarthi et al., 2022). Some of these remarks have the potential to stir up controversy on social media about a particular person or group of people. This demonstrates the need to prohibit the posting of these kinds of offensive comments on social media. Both the union territory of Puducherry (Pondicherry) and the Indian state of Tamil Nadu speak Tamil as their official language. In addition, it is spoken widely in Malaysia, Mauritius, Fiji, and South Africa. It is also the official language of Singapore and Sri Lanka.(Krishnamurti, 2022)

In the past, text classification was carried out on the text's sentence embedding using linear classifiers. This was followed by Recurrent Neural Networks. Transformers became prominent in the field of natural language processing following the publication (Vaswani et al., 2017). They have an attention layer mechanism that provides context to words in the text. The development of the transformer architecture has resulted in the development of numerous other transformer variations, such as BERT (Devlin et al., 2018), XLMRoBERTa (Conneau et al., 2019), MuRIL (Khanuja et al., 2021), etc.

In this paper, Transformer-based models and Recurrent Neural Network models are used for abusive comment detection in Tamil and code-mixed data(Tamil-English). The results from the performance of individual models and their ensemble are obtained to determine the best-performing model for this task.

2 Related works

Numerous published works have addressed the problem of identifying offensive content in high-resource languages. Umer et al. (2023)analysed the impact of FastText word embedding on text classification. Compared to contextual word embeddings, FastText has limited ability to capture complex se-

Third Workshop on Speech and Language Technologies for Dravidian Languages,
pages 160–165, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_021

mantic relationships. Fazil et al. (2023)demonstrated an attentional multi-channel convolutional-BiLSTM network for the classification of hateful content. GloVe embedding used in this paper may struggle with Out-Of-Vocabulary words. The pre-trained BERT model ensembled with Deep Learning (DL) models are the foundation of Mazari et al. (2023)proposed multi-aspect hate speech detection approach.This approach will have difficulty in handling multi-lingual or non-english data. Başarslan and Kayaalp (2023) proposed a deep learning model with multiple Bidirectional Gated Recurrent Units and Convolution layers for social media sentiment analysis.

Compared to English and other high-resource languages, low-resource languages have significantly fewer published studies on the detection of abusive comments. Datasets have been created by Chakravarthi et al. (2020) to promote research in Tamil, one of the Indian Dravidian languages.Rajalakshmi et al. (2023) proposed a method to detect hate speech or offensive content in Tamil. A detailed analysis was made to study the performance of stemming and pre-trained transformer models for word embedding. For the detection of offensive language in Tamil YouTube comments, Subramanian et al. (2022) proposed adapter-based transformer models. It was done using mBERT, MuRIL (Base and Large), and XLM-RoBERTa (Base and Large).Chakravarthi et al. (2023) proposed a novel approach of fusing MPNet with a deep neural network for detecting offensive language content in low-resource Dravidian languages

3 Dataset

The objective of this shared-task (Priyadharshini et al., 2023)is to determine whether a given comment contains abusive language. The annotations in the corpus are done at the comment or post level. Tamil and Tamil-English comments were gathered from the YouTube comment section for the Abusive Comment Detection Dataset (Priyadharshini et al., 2022). The dataset consists of a comment and its corresponding label from one of the nine labels in the dataset: Misandry, Counter-speech, Misogyny, Xenophobia, Hope-Speech, Homophobia, Transphobic, Not-Tamil, and None of-the-above. Only eight classes are classified because the 'Not-Tamil' class has no test or development data instances. A few weeks before the deadline for run submission, the testing dataset, which lacked labels, was made available. The labelled test dataset was made available by the organisers for verification purposes after the results were declared. The number of samples in the training, validation, and testing datasets for each class are listed in Tables 1 and 2.

Table 1: Class Wise Distribution in Training, Validation and Test dataset (Tamil)

Class	Train	Validation	Test
Misogyny	125	24	48
Misandry	446	104	127
Homophobia	35	8	8
Transphobic	6	2	2
Xenophobia	95	29	25
Hope-speech	86	11	26
Counter-speech	149	36	47
None-of-the-above	1296	346	416
Total	2238	560	699

Table 2: Class Wise Distribution of Training, Validation and Test dataset (Tamil-English)

Class	Train	Validation	Test
Misogyny	211	50	57
Misandry	830	218	292
Homophobia	172	43	56
Transphobic	157	40	58
Xenophobia	297	70	95
Hope-speech	213	53	70
Counter-speech	348	95	88
None-of-the-above	3720	919	1141
Total	5948	1488	1857

4 Methodology

The overall architecture for Abusive Comment Detection in Tamil and Tamil-English is given in Figure 1.

4.1 Pre-processing

For Tamil sub-task, Pre-processing is done to remove the noisy elements from the text.Usernames, URLs, Extra spaces and Emojis are removed. Transliteration is the process of changing a word's script while maintaining the sentence's semantic meaning and strictly adhering to the target language's syntactical structure. (Hande et al., 2021). The Tamil-English code-mixed dataset is transliterated to Tamil and combined with the existing Tamil dataset.

Table 1 makes clear that there is severe class

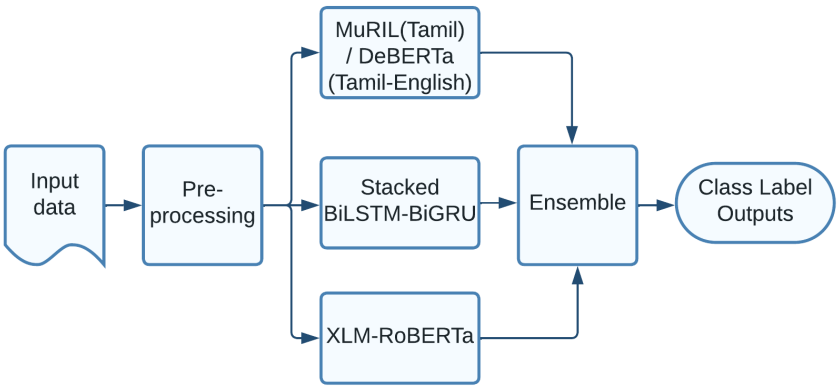

Figure 1: Abusive Comment Detection Architecture

imbalance in the dataset. An issue with class imbalance is when a dataset favours one class over another. By oversampling the minority classes, representation in the dataset is artificially increased, ensuring that the model receives sufficient exposure to learn patterns and make accurate predictions for those classes. Despite the potential benefits, stemming in Tamil NLP tasks is quite rare and hardly used. Tamil has a rich morphology, so the data was stemmed using a stemming algorithm created using the IndicNLP library's morphological analyzer (Kunchukuttan, 2022).

For Tamil-English sub-task, the dataset is pre-processed to remove noisy elements from the text.Usernames, URLs, Extra spaces and Emojis are removed. The pre-processed text is passed into the classification models.

4.2 Classification model

MuRIL (Khanuja et al., 2021) is a transformer based model built explicitly for Indian languages and trained on large amounts of Indic text corpora. XLM-RoBERTa is a multilingual variation of RoBERTa, which in and of itself outperformed BERT in a number of cutting-edge NLP tasks. 100 languages from 2.5 TB of filtered CommonCrawl data served as the basis for XLM-RoBERTa's pre-training (Conneau et al., 2019). DeBERTa improves RoBERTa and outperforms it in most NLP tasks by utilising enhanced mask decoding and disentangled attention (He et al., 2020).

MuRIL and XLM-RoBERTa pre-trained transformers are fine-tuned for Tamil dataset. For Tamil-English code-mixed dataset, DeBERTa and XLM-RoBERTa transformers are fine-tuned respectively. The fastText model is trained for a specified number of epochs(100) using the Skip-gram algorithm separately on Tamil and Tamil-English datasets.The word vectors generated after training are used for creating the embedding layer of the RNN model. The encoded sentences are then fed into two parallel, stacked recurrent layers, consisting of Bi-GRU and Bi-LSTM layers with the same number of units. The hidden states computed over all time steps from these recurrent layers are concatenated. The concatenated hidden states are then passed through global average pooling and global max pooling layers. The outputs of both pooling layers are concatenated again and served as the input for a dense classification layer. 'Adam's optimizer is used with a loss function of sparse_categorical_crossentropy. Table 3 represents the parameters used to finetune the MuRIL, XLM-RoBERTa, DeBERTa base models and the Bi-LSTM-Bi-GRU model.

The output class probabilities obtained from the transformer-based models and the stacked BiLSTM-BiGRU model are ensembled using weighted average method.MuRIL, XLM-RoBERTa and RNN model are ensembled for Tamil text. DeBERTa, XLM-RoBERTa and RNN model are ensembled for Tamil-English text. In the weighted averaging, Each model composing the ensemble model is assigned a weight depending on its individual performance. The values of individual weights range from zero to one , and the total sum of the weights given to the individual models is one.Equation 1 gives the formula for weighted average prediction.W_i: weight assigned to the i^{th} model.Y_i: prediction of the i^{th} model.

$$\hat{y} = \sum_{i=1}^{n} w_i \cdot y_i \quad (1)$$

Table 3: Hyper-parameters used for fine-tuning the models

	Tamil sub-task			Tamil-English sub-task		
Hyper-parameter	**MuRIL**	**XLM-RoBERTa**	**Bi-LSTM-Bi-GRU**	**DeBERTa**	**XLM-RoBERTa**	**Bi-LSTM-Bi-GRU**
Learning Rate	2e-5	2e-5	1e-3	2e-5	2e-5	1e-3
Batch Size	16	16	32	8	8	32
Number of Epochs	3	8	15	10	10	20

Table 4: Results for Abusive Comment Detection in Tamil Language (Validation dataset)

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
MuRIL	0.69	0.44	0.42	0.42
Bi-LSTM-Bi-GRU	0.66	0.31	0.37	0.32
XLM-RoBERTa	0.71	0.44	0.40	0.41
Ensemble	**0.74**	**0.50**	0.41	**0.43**

Table 5: Results for Abusive Comment Detection in Code-Mixed Language (Validation dataset)

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
DeBERTa	0.73	0.57	0.49	0.52
Bi-LSTM-Bi-GRU	0.74	0.45	0.60	0.50
XLM-RoBERTa	0.74	0.52	0.49	0.50
Ensemble	**0.78**	**0.71**	0.49	**0.56**

Table 6: Results for Abusive Comment Detection (Test datasets)

Dataset	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Tamil	0.69	0.40	0.45	0.41
Tamil-English	0.75	0.46	0.58	0.50

5 Results and Analysis

The training set was used to train all the models, and the development set was used to validate them.The organisers indicated macro F1-score as their primary evaluation metric.Apart from this,a few more performance metrics like accuracy and the macro averages of precision and recall are also used to assess the classification models. In order to determine the classification performance more accurately, all four evaluation metrics are calculated.The performance of the models on the Tamil and Tamil-English validation datasets is shown in Tables 4 and 5, respectively.On analysing the performance of all the models on validation datasets, it is clearly seen that the ensemble model performs better than the individual models. The weighted average ensemble model works better because it combines diverse representations, corrects errors, reduces bias, improves robustness to variability, and leverages the strengths of each model. By integrating multiple perspectives and leveraging their complementary abilities, the ensemble captures a broader range of patterns and linguistic nuances, leading to an improved performance. Due to higher macro-F1 score, the ensemble model is used in both cases for the test dataset. Table 6 contains the result obtained for Tamil and Tamil-English test datasets using ensemble model. The macro-F1 scores obtained using the ensemble model secured rank 2 for Tamil and rank 3 for Tamil-English in the shared task.

According to a comparison between predictions from the ensemble model for Tamil dataset and the original class labels of Tamil test data, the None-of-the-above class has the highest individual F1-score of 0.81 and Transphobic class has the lowest individual F1-score of 0 since it only has two data points in the test data. The None-of-the above class in the Tamil-English dataset has the highest individual F1-score of 0.85, while the Misogyny class has the lowest individual F1-score of 0.27.Table 7 contains the class-wise F1-score for both the datasets.

Table 7: Classwise F1-Score obtained using Ensemble Model on Test datasets

Class Name	Tamil	Tamil-English
Misogyny	0.42	0.27
Misandry	0.66	0.71
Homophobia	0.47	0.45
Transphobic	0.00	0.35
Xenophobia	0.28	0.65
Hope-Speech	0.30	0.33
Counter-speech	0.34	0.40
None-of-the-above	0.81	0.85

6 Conclusion and Future work

In this paper, a new approach has been proposed for abusive comment detection based on ensemble learning. The proposed model combined the pre-trained transformer models(MuRIL, XLM-RoBERTa, DeBERTa) with a deep-learning model built by stacking Bi-LSTMs and Bi-GRUs on FastText embeddings. This ensemble learning has actually reduced the number of misclassified instances and thus improved the precision of the abusive comment detection model. For Tamil dataset, an ensemble of MuRIL, Bi-LSTM-Bi-GRU and XLM-RoBERTa provided the best results with a macro-averaged F1 score of 0.41. For Tamil-English dataset, an ensemble of DeBERTa, Bi-LSTM-Bi-GRU and XLM-RoBERTa provided the best results with a macro-averaged F1 score of 0.50.

Further this work can be extended by exploring the performance of adapter-based transformer models. Ensembling of other transformer models with RNNs can also be explored in future.

References

Muhammet Sinan Başarslan and Fatih Kayaalp. 2023. Mbi-grumconv: A novel multi bi-gru and multi cnn-based deep learning model for social media sentiment analysis. *Journal of Cloud Computing*, 12(1):1–16.

Bharathi Raja Chakravarthi, Manoj Balaji Jagadeeshan, Vasanth Palanikumar, and Ruba Priyadharshini. 2023. Offensive language identification in dravidian languages using mpnet and cnn. *International Journal of Information Management Data Insights*, 3(1):100151.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John P McCrae. 2020. Corpus creation for sentiment analysis in code-mixed tamil-english text. *arXiv preprint arXiv:2006.00206*.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John Philip McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 2022. 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*, pages 369–377.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kayalvizhi Sampath, Durairaj Thenmozhi, Sathiyaraj Thangasamy, Rajendran Nallathambi, and John Phillip McCrae. 2021. Dataset for identification of homophobia and transophobia in multilingual youtube comments. *arXiv preprint arXiv:2109.00227*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mohd Fazil, Shakir Khan, Bader M Albahlal, Reemiah Muneer Alotaibi, Tamanna Siddiqui, and Mohd Asif Shah. 2023. Attentional multi-channel convolution with bidirectional lstm cell toward hate speech prediction. *IEEE Access*, 11:16801–16811.

Alon Halevy, Cristian Canton-Ferrer, Hao Ma, Umut Ozertem, Patrick Pantel, Marzieh Saeidi, Fabrizio Silvestri, and Ves Stoyanov. 2022. Preserving integrity in online social networks. *Communications of the ACM*, 65(2):92–98.

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. *arXiv preprint arXiv:2108.12177*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

B. Krishnamurti. 2022. Tamil language. encyclopedia britannica. https://www.britannica.com/topic/Tamil-language.

Prasanna Kumar Kumaresan, Premjith, 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 *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 16–18.

A. Kunchukuttan. 2022. The indicnlp library. https://github.com/anoopkunchukuttan/indic_nlp_library.

Ahmed Cherif Mazari, Nesrine Boudoukhani, and Abdelhamid Djeffal. 2023. Bert-based ensemble learning for multi-aspect hate speech detection. *Cluster Computing*, pages 1–15.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in tamil-acl 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Malliga S, SUBALALITHA CN, Kogilavani S V, Premjith B, Abirami Murugappan, and Prasanna Kumar Kumaresan. 2023. Overview of shared-task on abusive comment detection in tamil and telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

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.

Ratnavel Rajalakshmi, Srivarshan Selvaraj, Pavitra Vasudevan, et al. 2023. Hottest: Hate and offensive content identification in tamil using transformers and enhanced stemming. *Computer Speech & Language*, 78:101464.

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.

Benhur Sean. 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*, pages 279–285.

Malliga Subramanian, Rahul Ponnusamy, Sean Benhur, Kogilavani Shanmugavadivel, Adhithiya Ganesan, Deepti Ravi, Gowtham Krishnan Shanmugasundaram, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2022. Offensive language detection in tamil youtube comments by adapters and cross-domain knowledge transfer. *Computer Speech & Language*, 76:101404.

Muhammad Umer, Zainab Imtiaz, Muhammad Ahmad, Michele Nappi, Carlo Medaglia, Gyu Sang Choi, and Arif Mehmood. 2023. Impact of convolutional neural network and fasttext embedding on text classification. *Multimedia Tools and Applications*, 82(4):5569–5585.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.