Athena@DravidianLangTech: Abusive Comment Detection in Code-Mixed Languages using Machine Learning Techniques

Hema M, Anza Prem, Rajalakshmi S, Angel Deborah S
Department of Computer Science and Engineering,
Sri Sivasubramaniya Nadar College of Engineering, Chennai - 603110, Tamil Nadu, India
hema19037@cse.ssn.edu.in, anza19020@cse.ssn.edu.in
rajalakshmis@ssn.edu.in, angeldeborahs@ssn.edu.in

Abstract

The amount of digital material that is disseminated through various social media platforms has significantly increased in recent years. Online networks have gained popularity in recent years and have established themselves as go-to resources for news, information, and entertainment. Nevertheless, despite the many advantages of using online networks, mounting evidence indicates that an increasing number of malicious actors are taking advantage of these networks to spread poison and hurt other people. This work aims to detect abusive content in youtube comments written in the languages like Tamil, Tamil-English (code-mixed), Telugu-English (code-mixed). This work was undertaken as part of the ”DravidianLangTech@RANLP 2023” shared task. The Macro F1 values for the Tamil, Tamil-English, and Telugu-English datasets were 0.28, 0.37, and 0.6137 and secured 5^{th}, 7^{th}, 8^{th} rank respectively.

1 Introduction

The detection of abusive language has become a major area for research and development in computer linguistics and natural language processing (NLP). The requirement to recognise and filter out harmful or abusive content has multiplied enormously with the rise of Internet platforms and online social media. Detection of abusive or offensive content in English language has been done by many researchers in a wide manner [13, 14, 15]. Recent days, the interest goes on the low resource languages and native languages. In multilingual societies like India, where Tamil and Telugu are widely spoken, this is especially crucial.

Due to the complexity of their linguistic structures and the lack of sufficient labelled data for the training of strong models, the identification of abusive language presents particular difficulties in Tamil and Telugu. A wide range of harmful information is included in abusive language, such as hate speech, cyberbullying, vulgarity, and disparaging remarks. For the sake of upholding a secure and civilised online community, it is essential to accurately identify such offensive language in Tamil and Telugu.

The identification of abusive language in English has advanced significantly in recent years, according to academics. However, adapting these methods to Tamil and Telugu necessitates taking into account these languages’ unique traits, cultural quirks, and lack of labelled data. Furthermore, due to variances in syntax, morphology, and the existence of dialectical changes, existing models might not generalise effectively to Tamil and Telugu especially the deep learning techniques as most of them are built for English language.

This work aims to contribute to the development of efficient abusive language detection systems for Tamil and Telugu by tackling the difficulties unique to these languages, promoting a safer and more welcoming online environment for users who communicate in Tamil and Telugu.

In this study, multiple machine learning models were tested in an effort to create an effective system for identifying hate speech and abusive language in Tamil and Telugu comments on YouTube. The paper is organised as follows: Earlier studies on the identification of abusive language in Dravidian languages like Tamil and Telugu are discussed in Section 2. The proposed system along with the architecture diagram and the system modules are explained in Section 3. The datasets and methodologies used in the suggested system are presented in Section 4. The results are discussed in Section 5. The conclusion and future work are presented in Section 6.

Third Workshop on Speech and Language Technologies for Dravidian Languages, pages 147–151, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_019

2 Related Work

F. Balouchzahi et al. [3] centred on the detection of objectionable remarks in texts written in both the native script and Tamil written in code-mixed script. Two models were used to tackle this problem: n-gram-Multilayer Perceptron which makes use of an MLP classifier supplied with character-n-gram features, as well as the (1D ConvLSTM) model, were submitted. The n-gram MLP model performed better than the other two model, corresponding to weighted F1-scores of 0.430 for texts written in the native Tamil script and 0.560 for texts written in code-mixed Tamil, respectively.

Charangan Vasantharajan and Uthayasanker Thayasivam [2] offered a novel and flexible way of selective translation and transliteration operations to improve the results of adjusting and assembling BERT, DistilBERT, and XLM-RoBERTa. The experiment's findings proved that ULMFiT is the best model for the task. ULMFiT and mBERTBiLSTM beat other well-known transfer learning models like DistilBERT and XLM-RoBERTa as well as hybrid deep learning models for this Tamil code-mix dataset.

In the work done by Shantanu Patankar et al. [9] recurrent neural networks, ensemble models, and transformers were used to optimise the results. For the Tamil data, MuRIL and XLM-RoBERTA were utilised. The macro-averaged F1 score generated by the models was 0.43. With a macro-averaged f1 score of 0.45, the top models MuRIL and M-BERT both generated excellent results for the code-mixed data.

Malliga Subramanian et al. [5] worked to identify the offensive utterances, models based on traditional machine learning techniques—such as Bernoulli Naive Bayes, Support Vector Machine, Logistic Regression, and KNearest Neighbor—were constructed. The multilingual transformer-based pre-trained models of natural language processing mBERT, MuRIL (Base and Large), and XLM-RoBERTa (Base and Large) were also used in the experiments.

Pradeep Kumar Roy et al. [6] investigated the use of different machine learning and deep learning approaches. Combining the output of transformer and deep learning-based models, an ensemble model was proposed to detect hate speech and objectionable language on social networking sites. The experimental findings of the suggested weighted ensemble framework outperformed state-of-the-art models for the Malayalam and Tamil code-mixed datasets, achieving weighted F1-scores of 0.802 and 0.933, respectively.

3 Abusive Content Detection System

We have used the deep learning transformer model (BERT), machine learning models (Logistic Regression, Support Vector Machines, Decision Trees, Naive Bayes) and Ensemble model (Random Forest) for abusive content classification. The training dataset is used to build the model by learning the data, development dataset is used to evaluate and fine tune the trained model and test dataset is used for final prediction. The features are extracted from the input text and its matching label are used for learning and training. The texts in the dataset are vectorized using the Term Frequency - Inverse Document Frequency (TF-IDF) vectorization technique for feature extraction. The model that produces the greatest macro F1-Score is selected as the final model to be used for detection after various machine learning models are tested. The model is trained using the training dataset. The performance of the trained model is assessed using the development dataset. By tweaking the parameters, the model is re-trained based on the performance. Finally, predictions are made using the test data and the model. Figure 1 displays the architecture of the system.

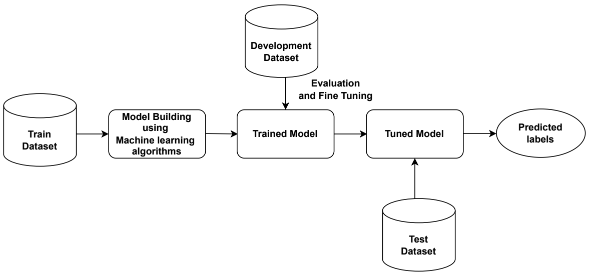

Figure 1: Architecture diagram

4 Implementation Modules

This section elaborates on the datatsets and the methodologies used.

4.1 Datasets Used

The abusive comment detection task consists of 3 datasets: Tamil, Tamil-English and Telugu-English [7, 8].

The **Tamil** dataset consisted of a total of 2240 sentences with their corresponding labels. It was a multi-class classification task which included the following labels: Misandry (446), Counter-speech (149), Misogyny (125), Xenophobia (95), Hope-Speech (86), Homophobia (35), Transphobic (6), Not-Tamil (2), None-of-the-above (1296).

The **Tamil-English** dataset consisted of a total of 5948 sentences with their corresponding labels. The labels present were Misandry (830), Counter-speech (348), Xenophobia (297), Hope-Speech (213), Misogyny (211), Homophobia (172), Transphobic (157), None-of-the-above (3720).

The **Telugu-English** dataset consisted of 4000 sentences along with their labels. This task was a binary classification task and included the labels: non-hate (2061) and hate (1939).

4.2 Methodology

Data Pre-processing: The text presented in the data cannot be passed to the model directly. Hence it is converted into numerical vectors by applying the TF-IDF vectorization on the text data present. After the vectorization the text data in the form of a dense matrix is passed to the model.

Training machine learning models: Logistic Regression (LR), Naive Bayes (NB), Support Vector Machines (SVM), Random Forest (RF), Decision Tree (DT) and BERT techniques are the machine learning models employed. The training dataset is used to train each model.

Evaluating machine learning models: The trained models are evaluated using the development dataset. The parameters of the model are fine-tuned based on the performance. The model that gives the maximum macro F1-Score is chosen as the final detection model.

Running model on test data: Results are obtained once the detection model receives the test data.

5 Results

Several machine learning models are experimented using the given datasets. Evaluation metrics like accuracy (Acc), precision (Prec), Recall and F1-score are used to evaluate the performance of the models. The model that gives the best accuracy and F1-score is chosen to run finally on the test data.

Table 1 displays the models' results on the Tamil dataset. It is evident that SVM provided the highest accuracy and F1-score. The models' results on the Tamil-English dataset are displayed in Table 2. As can be seen, SVM provides the highest level of accuracy and F1-score. Table 3 displays the models' results on the Telugu-English dataset. It is evident that Logistic Regression provides the highest level of accuracy and F1-score.

Model Name	**Acc**	**Prec**	**Recall**	**F1-Score**
LR	0.66	0.30	0.21	0.22
NB	0.67	0.27	0.16	0.16
SVM	0.68	0.36	0.29	0.26
RF	0.67	0.34	0.30	0.25
DT	0.54	0.23	0.24	0.23
BERT	0.63	0.13	0.16	0.14

Table 1: Comparison results for the Tamil training dataset

Model Name	**Acc**	**Prec**	**Recall**	**F1-Score**
LR	0.70	0.63	0.27	0.31
NB	0.69	0.42	0.20	0.21
SVM	0.72	0.69	0.33	0.38
RF	0.72	0.55	0.30	0.34
DT	0.65	0.39	0.36	0.37
BERT	0.71	0.35	0.25	0.27

Table 2: Comparison results for the Tamil-English training dataset

Model Name	**Acc**	**Prec**	**Recall**	**F1-Score**
LR	0.72	0.72	0.72	0.72
NB	0.70	0.70	0.70	0.70
SVM	0.69	0.69	0.69	0.69
RF	0.69	0.69	0.69	0.69
DT	0.65	0.65	0.65	0.65
BERT	0.68	0.71	0.69	0.68

Table 3: Comparison results for the Telugu-English training dataset

The task on the Tamil and the Tamil-English dataset are multi-class classification tasks. SVM has proved to work best on them due to the factors like effective separation of classes and resistance to overfitting.

The task on the Telugu-English dataset is a binary-class classification problem. LR has performed best on this dataset due to its simplicity, interpretability, efficiency with small dataset and robustness to outliers.

6 Performance Analysis

During testing, it was found that simpler machine learning models like SVM and LR performed better than more complex Transformer models like BERT and mBERT. This can be due to the dataset's quantity and quality. Traditional ML models generally require less data than more complex models to perform successfully. ML models might do better than BERT if the dataset isn't too big.BERT excels at a number of NLP tasks, however improved training and classification require a large dataset with a wide variety of texts. The performance of the BERT may suffer if the data are unstable or skewed.

It has been identified that even though the accuracy for all the three datasets (multi class and binary classification) are more or less same, the F1-score of multi-class classification models are very less when compared to the binary classification models. This is due to the imbalance in Tamil and Tamil-English dataset. We belive that this can be rectified using data augmentation in future.

We have also noticed that deep learning transformer model has not performed well as it could not learn the features effectively from the small dataset of low resource language. In order to improve the performance of the deep learning model we have planned to augment the dataset as a future work, so that it can solve the data imbalance as well as increase the number of samples in the dataset.

For the test dataset of Tamil, Tamil-English, Telugu-English languages we have achieved the F1-score of 0.28, 0.37, and 0.6137 respectively.

7 Conclusion and Future Work

This task was taken as a part of "Abusive Comment Detection in Tamil and Telugu at DravidianLangTech@RANLP 2023" shared task. For the Tamil and Tamil-English datasets (multi class classification) Support Vector Machine showed the maximum performance and hence was used to run on the test data and the results were submitted. For the Telugu-English dataset (binary classification) Logistic Regression showed the best performance hence that was used to run on the test data. Our team "Athena" was ranked 5^{th}, 7^{th}, 8^{th} for the Tamil, Tamil-English and Telugu-English datasets respectively.

In the future we would like to improve the results by using larger and more balanced datasets. We would also like to experiment on vectorization feature extraction techniques more suitable for code-mixed and non-English languages.

References

[1] Adaikkan Kalaivani, Durairaj Thenmozhi and Chandrabose Aravindan, TOLD: Tamil Offensive Language Detection in Code-Mixed Social Media Comments using MBERT with Features based Selection, CEUR Workshop Proceedings (2021)

[2] Charangan Vasantharajan, Uthayasanker Thayasivam, Towards Offensive Language Identification for Tamil Code-Mixed YouTube Comments and Posts, arXiv:2108.10939 [cs.CL] (2021)

[3] F. Balouchzahi, M. D. Anusha, H. L. Shashirekha, G. Sidorov, MUCIC@TamilNLP-ACL2022: Abusive Comment Detection in Tamil Language using 1D Conv-LSTM, Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, ACL (2022)

[4] Gayathri G L, Krithika S, Divyasri K, Durairaj Thenmozhi, B. Bharathi, PANDAS@TamilNLP-ACL2022: Abusive Comment Detection in Tamil Code-Mixed Data Using Custom Embeddings with LaBSE, Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, ACL (2022)

[5] Malliga Subramanian, Rahul Ponnusamy, Sean Benhur, Kogilavani Shanmugavadivel, Adhithiya Ganesan, Deepti Ravi, Gowtham Krishnan Shanmugasundaram, Ruba Priyadharshini, Bharathi Raja Chakravarthi, Offensive language detection in Tamil YouTube comments by adapters and cross-domain knowledge transfer, Elsevier, Computer Speech & Language (2022)

[6] Pradeep Kumar Roy, Snehaan Bhawal, Chinnaudayar Navaneethakrishnan Subalalitha, Hate speech and offensive language detection in Dravidian languages using deep ensemble framework, Elsevier, Computer Speech & Language (2022)

[7] Priyadharshini, Ruba and Chakravarthi, Bharathi Raja and Chinnaudayar Navaneethakrishnan, Subalalitha and Subramanian, Malliga and Shanmugavadivel, Kogilavani and B, Premjith and Murugappan, Abirami and Karnati, Sai Prashanth and Rishith and Janakiram, Chandu and Kumaresan, Prasanna Kumar, Findings of the Shared Task on Abusive Comment Detection in Tamil and Telugu, Recent Advances in Natural Language Processing (2023)

[8] Priyadharshini, Ruba and Chakravarthi, Bharathi Raja and Cn, Subalalitha and Durairaj, Thenmozhi and Subramanian, Malliga and Shanmugavadivel, Kogilavani and U Hegde, Siddhanth and Kumaresan, Prasanna, Overview of Abusive Comment Detection in Tamil-ACL 2022, Association for Computational Linguistics (2022)

[9] Shantanu Patankar, Omkar Gokhale, Onkar Litake, Aditya Mandke, Dipali Kadam, Optimize-Prime@DravidianLangTech-ACL2022: Abusive Comment Detection in Tamil, arXiv:2204.09675 [cs.CL] (2022)

[10] Snehaan Bhawal, Pradeep Kumar Roy, Abhinav Kumar, Hate Speech and Offensive Language Identification on Multilingual code-mixed Text using BERT, CEUR Workshop Proceedings, FIRE (2021)

[11] Varsha Pathaka, Manish Joshib, Prasad Joshic, Monica Mundadad, Tanmay Joshie, KBCNMUJAL@HASOC-Dravidian-CodeMixFIRE2020: Using Machine Learning for Detection of Hate Speech and Offensive Code-Mixed Social Media text, CEUR Workshop Proceedings, FIRE (2020)

[12] Vikram Gupta, Sumegh Roychowdhury, Mithun Das, Somnath Banerjee, Punyajoy Saha, Binny Mathew, Hastagiri Vanchinathan, Animesh Mukherjee, MACD: Multilingual Abusive Comment Detection at Scale for Indic Languages, 36th Conference on Neural Information Processing Systems, NeurIPS (2022)

[13] Mosquera, Alejandro. "amsqr at SemEval-2020 Task 12: Offensive language detection using neural networks and anti-adversarial features." In Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 1898-1905. 2020.

[14] Sivanaiah, Rajalakshmi, Angel Suseelan, S. Milton Rajendram, and Mirnalinee Tt. "TECHSSN at SemEval-2020 Task 12: Offensive language detection using BERT embeddings." In Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 2190-2196. 2020.

[15] Sivanaiah, Rajalakshmi, S. Milton Rajendram, Mirnalinee Tt, Abrit Pal Singh, Aviansh Gupta, and Ayush Nanda. "Techssn at semeval-2021 task 7: Humor and offense detection and classification using colbert embeddings." In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pp. 1185-1189. 2021.