AbhiPaw@ DravidianLangTech: Abusive Comment Detection in Tamil and Telugu using Logistic Regression

Abhinaba Bala
IIIT Hyderabad, India
abhinaba.bala@research.iiit.ac.in

Parameswari Krishnamurthy
IIIT Hyderabad, India
param.krishna@iiit.ac.in

Abstract

Abusive comments in online platforms have become a significant concern, necessitating the development of effective detection systems. However, limited work has been done in low resource languages, including Dravidian languages. This paper addresses this gap by focusing on abusive comment detection in a dataset containing Tamil, Tamil-English and Telugu-English code-mixed comments. Our methodology involves logistic regression and explores suitablef embeddings to enhance the performance of the detection model. Through rigorous experimentation, we identify the most effective combination of logistic regression and embeddings. The results demonstrate the performance of our proposed model, which contributes to the development of robust abusive comment detection systems in low resource language settings.

Keywords: Abusive comment detection, Dravidian languages, logistic regression, embeddings, low resource languages, code-mixed dataset.

1 Introduction

The widespread prevalence of abusive comments in online social networks (OSNs) has raised serious concerns about the safety and well-being of users. Detecting and mitigating abusive content has become a paramount objective in fostering a respectful and inclusive online environment. While significant progress has been made in abusive comment detection, much of the existing research primarily focuses on high-resource languages, leaving a critical gap in addressing this issue in low resource languages. The lack of resources and research devoted to comprehending and addressing abusive content in these languages poses a hindrance to the development of efficient detection systems.

In recent studies conducted by Abusive Comment Detection in Tamil - ACL 2022 (Priyadharshini et al., 2022), a shared task was introduced to detect categories of abusive comments in social media, focusing on languages such as Tamil and code-mixed language containing Tamil and English scripts. Swaminathan et al. (Swaminathan et al., 2022) proposed a classification model that combines language-agnostic sentence embeddings with TF-IDF vector representation, employing traditional classifiers. Balouchzahi et al. (Balouchzahi et al., 2022) addressed abusive comment detection in native Tamil script texts and code-mixed Tamil texts using n-gram-Multilayer Perceptron (n-gram-MLP) and 1D Convolutional Long Short-Term Memory (1D Conv-LSTM) models. S.N. et al. (S N et al., 2022) employed TF-IDF with char-wb analyzers and the Support Vector Machine (SVM) classifier with a polynomial kernel for classification. These studies demonstrate initial attempts to tackle abusive comment detection in low resource languages, but further research is necessary to address the challenges specific to these languages comprehensively.

The unique linguistic characteristics and nuances of Dravidian languages present significant challenges in detecting abusive comments accurately. Additionally, the limited availability of annotated datasets and language-specific linguistic features further exacerbate these challenges. Therefore, it is crucial to explore novel approaches and methodologies tailored specifically for detecting abusive comments in Dravidian and other low resource languages.

In this paper, we aim to bridge the research gap by comprehensively reviewing the state-of-the-art techniques and methodologies employed in abusive comment detection. Our primary focus will be on addressing the challenges faced in low resource languages, with a specific emphasis on Dravidian languages.

The outcomes of this research have significant

Third Workshop on Speech and Language Technologies for Dravidian Languages, pages 231–234, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_033

implications, as they contribute to building safer online communities in low resource language contexts, empower content moderators, and facilitate the development of automated systems capable of detecting and mitigating abusive comments effectively. By addressing the specific needs of Dravidian languages, we pave the way for further research in other low resource languages, promoting a more inclusive and equitable digital space.

In summary, this paper serves as a significant step towards understanding and combating abusive comment detection in low resource languages, with a particular focus on Dravidian languages. By leveraging innovative techniques and proposing tailored solutions, we aim to make substantial progress in creating a safer and more respectful online environment for users of diverse linguistic backgrounds. This work was submitted as a part of the DravidianLangTech workshop, 2023 (Priyadharshini et al., 2023b).

2 Related Work

Numerous studies have been conducted to identify abusive comments in various languages; however, there has been relatively less work done in low resource languages, highlighting a research gap in this area.

(Priyadharshini et al., 2023a), (Priyadharshini et al., 2022) conducted a shared task (at ACL 2022) that aims at detecting the categories of abusive comments that are posted on social media. They aggregate the comments from social media in two languages, namely, Tamil and in code mixed language containing Tamil and English scripts.

(Swaminathan et al., 2022) approached by building a classification model which includes different methods of feature extraction and the use of traditional classifiers. They propose a novel method of combining language-agnostic sentence embeddings with the TF-IDF vector representation that uses a curated corpus of words as vocabulary, to create a custom embedding, which is then passed to an SVM classifier. (Balouchzahi et al., 2022) addresses the abusive comment detection in native Tamil script texts and code-mixed Tamil texts. To address this challenge, two models: i) n-gram-Multilayer Perceptron (n-gram-MLP) model utilizing MLP classifier fed with char-n gram features and ii) 1D Convolutional Long Short-Term Memory (1D Conv-LSTM) model, were submitted. (S N et al., 2022) used TF-IDF with char-wb analyzers with Random Kitchen Sink (RKS) algorithm to create feature vectors and the Support Vector Machine (SVM) classifier with polynomial kernel for classification.

Transformer is an attention-based technique that effectively captures the contextual connections between words (or subwords) within a text, attracting significant attention in the field. (García-Díaz et al., 2022) sbow a knowledge integration strategy that combines sentence embeddings from BERT, RoBERTa, FastText and a subset of language-independent linguistic features. (Prasad et al., 2022) present XLM-RoBERTa and DeBERTa models for two multi-class text classification tasks in Tamil. (Biradar and Saumya, 2022) used a pre-trained transformer model such as Indic-bert for feature extraction, and on top of that, SVM classifier is used for stance detection. (Pahwa, 2022) presented an exploration of different techniques which can be used to tackle and increase the accuracy of models using data augmentation in NLP. (B and Varsha, 2022) used pre-trained transformer models such as BERT,m-BERT, and XLNET.

Certain works experiment with multiple methods individually to find the best performing model. (Rajalakshmi et al., 2022) approached the task with three methodologies - Machine Learning, Deep Learning and Transformer-based modeling. For Machine Learning, eight algorithms were implemented, among which Random Forest gave the best result with Tamil+English dataset for Deep Learning, Bi-Directional LSTM gave best result with pre-trained word embeddings. In Transformer-based modeling, they used IndicBERT and mBERT with fine-tuning, among which mBERT gave the best result. (Hossain et al., 2022) employed three machine learning (LR, DT, SVM), two deep learning (CNN+BiLSTM, CNN+BiLSTM with FastText) and a transformer-based model (Indic-BERT). The experimental results show that Logistic regression (LR) and CNN+BiLSTM models outperformed the others. (Bhattacharyya, 2022) experimented with logistic regression, SVMs, gradient boost classifier, finetuned MuRIL; with gradient boost classifier emerged to be the best performing method. (Patankar et al., 2022) also used three methods to optimize their results: Ensemble models, Recurrent Neural Networks, and Transformers.

3 Method

The objective of this task is to determine if a given YouTube comment exhibits abusive content.

3.1 Feature Extraction

The TfidfVectorizer is a feature extraction technique commonly used in natural language processing (NLP) and text mining tasks. ”TF-IDF” stands for Term Frequency-Inverse Document Frequency, which represents the importance of a word within a document in a corpus [cite :: (Sammut and Webb, 2010)].

The TfidfVectorizer calculates the TF-IDF value for each term (word) in a document by considering its frequency within the document and the inverse document frequency across the entire corpus. The TF-IDF value reflects the significance of a term in a document relative to its frequency in other documents.

The TfidfVectorizer converts a collection of raw text documents into a matrix representation where each row corresponds to a document and each column corresponds to a term. The matrix elements represent the TF-IDF values of the terms in the documents.

This vectorization technique is widely used for tasks such as text classification, information retrieval, and document similarity analysis. It helps capture the discriminative and important terms in a document while downweighting common and less informative terms.

3.2 Classifiers

The Logistic Regression model, trained on the feature vectors, employs a multi-class variant of the algorithm to classify comments into multiple abusive categories. It estimates the probability of each class and assigns the comment to the category with the highest probability. The model's training objective involves optimizing the parameters to minimize the discrepancy between the predicted probabilities and the true class labels across all classes.

4 Experiments and Results

4.1 Evaluation Metrics

The macro average F1-score is the performance metric used to evaluate the overall effectiveness of the detection model. It is derived by calculating the F1-score for each individual class and then taking the average across all classes. Regardless of class size or class imbalance, the macro average F1-score considers the performance of each class independently and then computes the average, giving equal importance to all classes.

4.2 Datasets

The dataset is provided by (Priyadharshini et al., 2023a), (Priyadharshini et al., 2022) as a shared task challenge for Abusive Comment Detection in Tamil and Telugu-DravidianLangTech@RANLP 2023. There are three datasets - Tamil, Tamil-English and Telugu-English

4.3 Results

We achieved a macro F1-score a. 0.27 in Tamil, 0.29 in Tamil-English, 0.6319 in Telugu languages. The current scores are suboptimal, indicating room for enhancement.

5 Conclusion

By transforming textual data into numerical features and employing the logistic function for multi-class classification, Logistic Regression enables accurate identification and categorization of abusive comments into multiple abusive categories, facilitating comprehensive analysis and understanding of the data.

As future work we look forward to use fine tuning on the specific datasets and investigate cross-lingual transfer learning.

References

Bharathi B and Josephine Varsha. 2022. SSNCSE NLP@TamilNLP-ACL2022: Transformer based approach for detection of abusive comment for Tamil language. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 158–164, Dublin, Ireland. Association for Computational Linguistics.

Fazlourrahman Balouchzahi, Anusha Gowda, Hosahalli Shashirekha, and Grigori Sidorov. 2022. MUCIC@TamilNLP-ACL2022: Abusive comment detection in Tamil language using 1D conv-LSTM. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 64–69, Dublin, Ireland. Association for Computational Linguistics.

Aanisha Bhattacharyya. 2022. Aanisha@TamilNLP-ACL2022:abusive detection in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 214–220, Dublin, Ireland. Association for Computational Linguistics.

Shankar Biradar and Sunil Saumya. 2022. IIITDWD@TamilNLP-ACL2022: Transformer-based approach to classify abusive content in Dravidian code-mixed text. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 100–104, Dublin, Ireland. Association for Computational Linguistics.

José García-Díaz, Manuel Valencia-Garcia, and Rafael Valencia-García. 2022. UMUTeam@TamilNLP-ACL2022: Abusive detection in Tamil using linguistic features and transformers. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 45–50, Dublin, Ireland. Association for Computational Linguistics.

Alamgir Hossain, Mahathir Bishal, Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. COMBATANT@TamilNLP-ACL2022: Fine-grained categorization of abusive comments using logistic regression. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 221–228, Dublin, Ireland. Association for Computational Linguistics.

Bhavish Pahwa. 2022. BpHigh@TamilNLP-ACL2022: Effects of data augmentation on indic-transformer based classifier for abusive comments detection in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 138–144, Dublin, Ireland. Association for Computational Linguistics.

Shantanu Patankar, Omkar Gokhale, Onkar Litake, Aditya Mandke, and Dipali Kadam. 2022. Optimize_Prime@DravidianLangTech-ACL2022: Abusive comment detection in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 235–239, Dublin, Ireland. Association for Computational Linguistics.

Gaurang Prasad, Janvi Prasad, and Gunavathi C. 2022. GJG@TamilNLP-ACL2022: Using transformers for abusive comment classification in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 93–99, Dublin, Ireland. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and Prasanna Kumar Kumaresan. 2023a. Findings of the shared task on Abusive Comment Detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages DravidianLangTech 2023*. Recent Advances in Natural Language Processing.

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, Dublin, Ireland. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Malliga S, SUBALALITHA CN, Kogilavani S V, Premjith B, Abirami Murugappan, and Prasanna Kumar Kumaresan. 2023b. 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.

Ratnavel Rajalakshmi, Ankita Duraphe, and Antonette Shibani. 2022. DLRG@DravidianLangTech-ACL2022: Abusive comment detection in Tamil using multilingual transformer models. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 207–213, Dublin, Ireland. Association for Computational Linguistics.

Prasanth S N, R Aswin Raj, Adhithan P, Premjith B, and Soman Kp. 2022. CEN-Tamil@DravidianLangTech-ACL2022: Abusive comment detection in Tamil using TF-IDF and random kitchen sink algorithm. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 70–74, Dublin, Ireland. Association for Computational Linguistics.

Krithika Swaminathan, Divyasri K, Gayathri G L, Thenmozhi Durairaj, and Bharathi B. 2022. PANDAS@abusive comment detection in Tamil code-mixed data using custom embeddings with LaBSE. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 112–119, Dublin, Ireland. Association for Computational Linguistics.