Multilingual Pretrained Models for Offensive Language Identification in Dravidian Code-Mixed Text

Asha Hegde[1], Amrithkala M Shetty[2], Shazia Mannan[3] and Sharal Coelho[1]

[1]Department of Computer Science, Mangalore University, India
[2]Department of Computer Applications, Nitte Institute of Professional Education, Nitte (Deemed to be University), Karnataka, India
[3]Department of Computer Science, Yenepoya (Deemed to be University), Bangalore Campus, Karnataka, India

Abstract

Offensive language identification in countries like India poses several challenges due to the usage of code-mixed and romanized variants of multiple languages by the users in their posts on social media. The challenge of offensive language identification on social media for Dravidian languages is harder, considering the low resources available for the same. In this paper, we employed two multilingual transformer-based language models–multilingual BERT (mBERT) and XLM-RoBERTa–to perform offensive language identification in Dravidian code-mixed languages. The languages in the task are romanized variants of four Dravidian languages: Malayalam, Tamil, Tulu, and Kannada. The proposed methodology makes use of Transfer Learning (TL) based Multilingual Bidirectional Encoder Representations from Transformers (mBERT) model and XLM-RoBERTa. These models achieved macro F1 scores of 0.475, 0.774, 0.453, and 0.819, securing 1st, 1st, 1st, and 2nd ranks for Kannada, Tamil, Tulu, and Malayalam languages, respectively. The proposed models yielded good results and are promising for effective offensive speech identification in low resource settings.

Keywords

Multilingual BERT, XLM-RoBERTa, code-mixed, Transfer Learning

1. Introduction

Offensive speech is defined as speech that causes a person to feel upset, resentful, annoyed, or insulted. In recent years, social media such as Twitter, Facebook and Reddit have been increasingly used for the propagation of offensive speech and the organization of hate and offense-based activities [1]. In a country like India with multiple native languages, users prefer to use their regional language in their social media interactions. It has also been identified that users tend to use roman characters for texting instead of the native script. This poses a severe challenge for the identification of offensive speech, considering the under-developed methodologies for handling code-mixed and romanized text.

Until a few years ago, hate and offensive speech were identified manually which is now an impossible task due to the enormous amounts of data being generated daily on social media platforms. The need for scalable automated methods for detecting hate speech has attracted significant research in the domains of natural language processing and Machine Learning. Researchers develop and experiment with a variety of techniques and tools such as bag of words models, N-grams, dictionary-based approaches, and word sense disambiguation techniques. Recent developments in multilingual text classification are led by Transformer architectures like mBERT and and XLM-RoBERTa.

An additional advantage of these architectures, particularly XLM-RoBERTa, is that they give good results even with lower-resource languages, and this particular aspect is beneficial to Indian languages, which do not have properly established datasets. In our work, we focused on using these architectures in multiple ways. However, there is a limitation in directly applying these models to romanized or code-mixed text. The transformer models are trained on languages in their native script, not in the romanized script in which users prefer to write online.

This shared task presents a corpus for offensive language identification of code-mixed text in Dravidian languages (Tamil-English, Malayalam-English, Kannada-English, and Tulu-English). This task is further

Forum for Information Retrieval Evaluation, December 17-20, 2025, India

hegdekasha@gmail.com (A. Hegde); amrithkalas@gmail.com (A. M. Shetty); shaziamannan.blr@yenepoya.edu.in (S. Mannan); sharalmucs@gmail.com (S. Coelho)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-02-24

complicated in low-resource languages where limited annotated datasets exist for offensive speech detection.

The paper is organized as follows: Section 2 explores related work, followed by an overview of the datasets in Section 3. Section 4 details the methodology, and Section 5 analyzes the results. The paper wraps up with conclusions in Section 6.

2. Related Work

Researchers used a wide variety of techniques for the identification of offensive languages. Few of the research studies are described below.

Mishra et al.[2] fine-tuned monolingual and multilingual BERT based network models to achieve better results in identifying hate speech. Mishra et al.[2] fine-tuned monolingual and multilingual BERT-based network models to gain more reasonable results in identifying hate speech. Code-mixed Sentiment Analysis (SA) has gained attention due to the increasing use of multilingual text on content social media platforms. For Tulu, a low-resource Dravidian language, Hegde et al. [3] created a gold-standard annotated corpus of 7,171 YouTube comments in trilingual code-mixed form (Tulu-Kannada-English), filling a major gap in resources. Their experiments using baseline ML models on this dataset demonstrated encouraging results, laying the groundwork for future progress in Tulu SA [4].

In [5], the paper addresses the identification of offensive speech in code-mixed and romanized Dravidian languages (Malayalam, Tamil, and Kannada), a challenging task due to low-resources settings. The study explores zero-shot and few-shot learning with multilingual models, primarily XLM-RoBERTa and mBERT[1], enhanced through selective translation, transliteration, fine-tuning, and ensembling. Additionally, interlanguage, inter-task, and multi-task transfer learning strategies are employed to use English resources. The proposed approaches achieve strong results, demonstrating their effectiveness for offensive speech detection in low-resource multilingual contexts. The work [6] proposes an ensemble deep learning model combining CNN and DNN for hate speech and offensive post detection in Malayalam and Tamil code-mixed and script-mixed MIoT texts. Using word-level, character-level, and Term Frequency-Inverse Document Frequency features, the model achieved better performance with weighted F1-scores of 0.91 (Tamil code-mixed), 0.78 (Malayalam code-mixed), and 0.95 (Malayalam script-mixed).

The work [7] introduces a multilingual, manually annotated dataset of over 60,000 YouTube comments in Tamil-English (44k), Kannada-English (7k), and Malayalam-English (20k) for SA and offensive language identification. The dataset, covering diverse code-mixing phenomena, was annotated by volunteers with high inter-annotator agreement. The paper [8] proposes a multi-head attention-based model for detecting offensive content in code-mixed Tamil social media comments. Features extracted from the attention model are classified using multiple algorithms, and final predictions are refined through majority voting. Evaluated on the HASOC 2021 Tamil CodeMix dataset (Task 2–Subtask 1), the system achieved a weighted F1 score of 0.83, securing 3rd place in the official rankings.

Limited research has been conducted on offensive language identification for Dravidian languages to date. This work addresses the existing research gap by developing methods for identifying offensive speech in Dravidian languages. The proposed systems are designed to be adaptable, with potential extensions to other Indian and international languages.

3. Task Description

This shared task presents a gold-standard dataset for offensive language detection in Tamil, Malayalam, Kannada, and Tulu, enabling researchers to develop robust classification models. The more details of the shared task are displayed in the task website [2]. The dataset consists of social media comments and posts that are categorized into four classes:

[1] http://huggingface.co/google-bert/bert-base-multilingual-cased

[2] https://dravidian-codemix.github.io/2025/

- Not Offensive (NO): Content without any offensive elements.
- Offensive Untargeted (OU): Offensive content that is not directed at a specific individual or entity.
- Offensive Targeted (OT): Direct attacks on an individual or group, including hate speech targeting a community, ethnicity, caste, or gender.
- Not Tamil/Not Malayalam/Not Kannada/Not Tulu (NT): Content that does not contain the Tamil, Malayalam, Kannada, or Tulu languages.

The primary goal of this shared task is to build and evaluate systems that can automatically classify social media text into these four categories. Participants will be provided with training, development, and test datasets to develop their models. Given the real-world class imbalance in offensive content, models must be designed to handle the skewed distribution of data effectively.

4. Methodology

In this study, we employed two multilingual transformer-based language models–multilingual BERT (mBERT) and XLM-RoBERTa–to perform multilingual text classification tasks. These models are selected due to their proven efficacy in handling diverse languages and contextual embeddings. To address potential class imbalance in the dataset, we experimented with class weighting techniques during training, which adjust the loss function to penalize misclassifications of underrepresented classes more heavily. Specifically, class weights were computed inversely proportional to the class frequencies in the training data.

- **mBERT with class weights** : This setup utilized the base mBERT model (bert-base-multilingual-cased) fine-tuned with class weights applied to the cross-entropy loss. This configuration yielded the best overall performance, achieving superior accuracy, F1-score, and generalization across languages, likely due to mBERT's balanced multilingual pre-training combined with effective handling of class imbalance.
- **XLM-RoBERTa with class weights**: We fine-tuned the base XLM-RoBERTa model (xlm-roberta-base) incorporating the same class weighting strategy. While it demonstrated strong results, particularly in cross-lingual transfer, it underperformed compared to the mBERT variant with weights, possibly owing to differences in pre-training objectives (e.g., XLM-RoBERTa's focus on masked language modeling across more languages but with less emphasis on next-sentence prediction).
- **mBERT without class weights**: This served as a baseline, fine-tuning mBERT without any class weighting, relying solely on standard cross-entropy loss. This approach resulted in noticeably different predictions, with a bias toward majority classes and reduced performance on minority classes, highlighting the importance of class weights in mitigating imbalance issues.

These models are fine-tuned for offensive language detection. This involves training the model on task-specific labeled data. During both pretraining and fine-tuning, the model utilizes attention mechanisms to process the input text. It considers the context of each word by attending to its surrounding words, capturing long-range dependencies effectively. Thus, mBERTmodel is designed to provide a powerful and flexible framework for multilingual NLP tasks such as SA[4], named entity recognition, and language identification, utilizing its pre-trained knowledge and ability to handle code-mixed text effectively with the multilingual support.

5. Experiments and Results

Statistics of the dataset provided by the organizers of the shared task for the Offensive Language Identification in Dravidian Languages for Tasks are shown in Tables 1. bert-base-multilingual-cased- a mBERT model from the huggingface repository is used to extract the feature vectors. After loading

Table 1
Languages - Dataset Statistics

Language	Train set	Development set	Test set	Total
Tamil	35,139	4,388	4,392	43,919
Malayalam	16,010	1,999	2,001	20,010
Kannada	6,217	777	778	7,772
Tulu	2,692	577	576	3,845

Table 2
Results of the proposed models with respect to different datasets

Language	Accuracy	Macro Precision	Macro Recall	Macro F1	Weighted Precision	Weighted Recall	Weighted F1	Ranks
Kannada	0.724	0.539	0.447	0.475	0.699	0.724	0.704	**1**
Malayalam	0.835	0.832	0.814	0.819	0.838	0.835	0.833	**2**
Tamil	0.964	0.859	0.72	0.774	0.962	0.964	0.962	**1**
Tulu	0.758	0.506	0.427	0.453	0.735	0.758	0.742	**1**

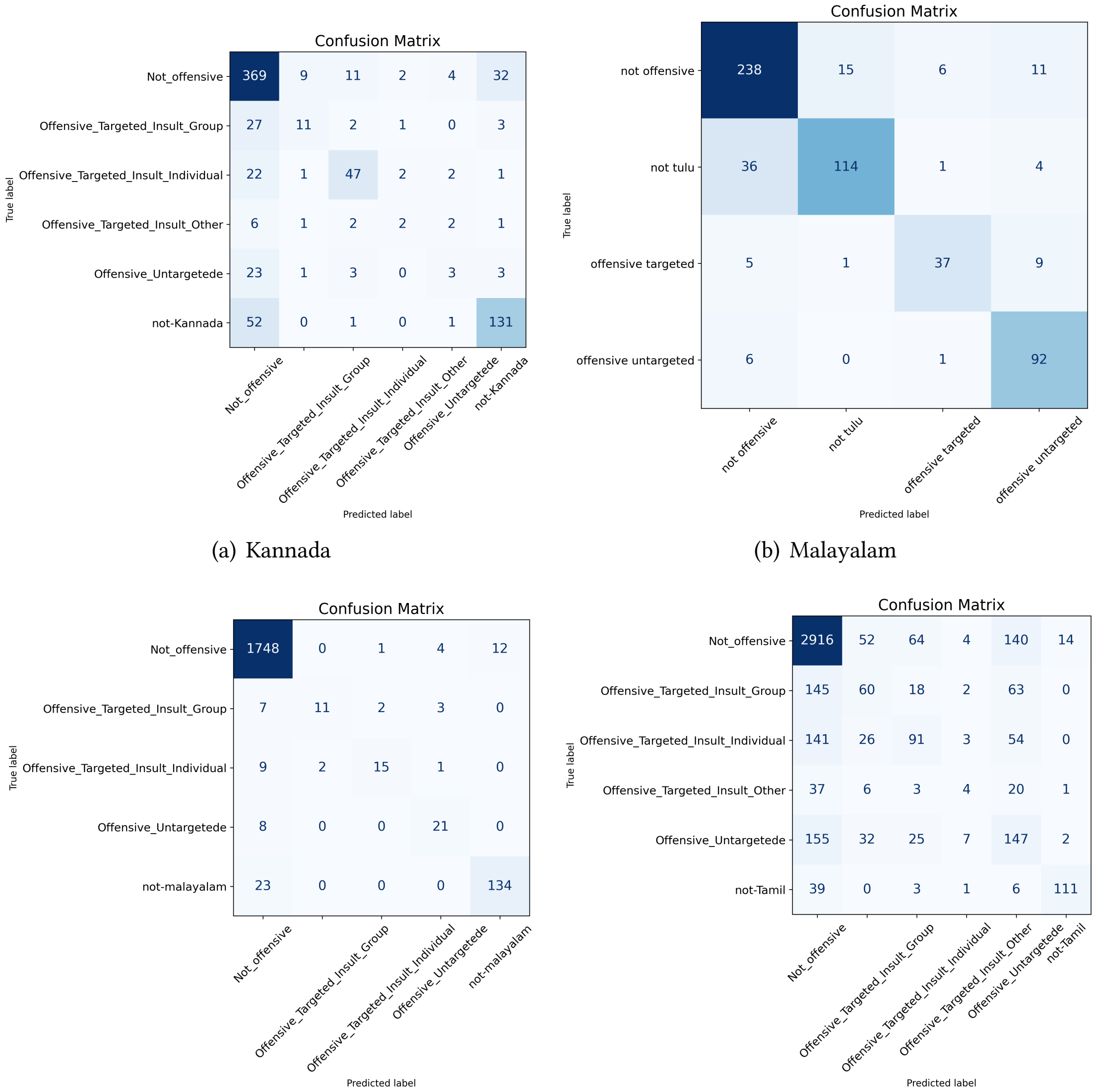
Table 3
Samples of misclassification Tulu and Kannada language datasets

Language	Comment	Actual Label	Predicted Label	Remarks
Tulu	Super. But minimize Kannada word use real tulu word in song	not offensive	not tulu	As the words are not Tulu our model predicted it as not Tulu
	guru kiran gud commented	not tulu	not offensive	As the comment consist ’gud’ represents ’Good’, it predicted as not offensive
Kannada	Bgm is good	Not_offensive	not-Kannada	Though it is not_offensive, also it belongs to not Kannada sentence
	super sir	not-Kannada	Not_offensive	Though it is not-Kannada it is also belongs to ’not_offensive’

the pre-trained mBERT model with its default parameter values, the model is frozen to prevent further updates to its weights.

The models are evaluated based on weighted F1 scores by incorporating class weights. Among the proposed models, mbert model with weights obtained better results. The proposed models obtained macro F1 scores of 0.475, 0.774, 0.453, and 0.819, securing 1st, 1st, 1st, and 2nd ranks for Kannada, Tamil, Tulu, and Malayalam languages, respectively.

From the Table 2, it is clear that the submitted predictions has exhibited better weighted F1 scores comparing to other teams by securing top ranks. The proposed methodology has still exhibited low macro F1 score for Kannada and Tamil text. The performance of the classifier across all four languages is illustrated using confusion matrices in Figure 1. These matrices highlight not only the correctly classified samples along the diagonal, but also the types of misclassifications that occur between classes. It can be observed that most of the wrong classifications are due to lack of context. Table 3 shows the sample text from Tulu and Kannada labeled Test sets along with their the actual and predicted labels. The proposed model achieved lower macro F1 scores for Kannada and Tulu compared to the other two languages evaluated. This reduced performance can be attributed to the significantly smaller dataset sizes available for Kannada and Tulu as shown in Table 1. The limited training data likely hindered the model’s ability to actually capture the linguistic nuances and code-mixed patterns prevalent in these low-resource languages.



(a) Kannada

(b) Malayalam

(c) Tamil

(d) Tulu

Figure 1: Confusion Matrices for mBERT Model Performance on Offensive Language Identification Across Four Languages

6. Conclusion

This paper describes the models submitted to the shared task ”Offensive Language Identification in Dravidian Languages” at FIRE 2025 [9]. The proposed methodology makes use of Transfer Learning (TL) based Multilingual Bidirectional Encoder Representations from Transformers (mBERT) model and XLM-RoBERTa. These models achieved macro F1 scores of 0.475, 0.774, 0.453, and 0.819, securing 1st, 1st, 1st, and 2nd ranks for Kannada, Tamil, Tulu, and Malayalam languages, respectively. The proposed models yielded good results and are promising for effective offensive speech identification in low resource settings.

Declaration on Generative AI

During the preparation of this work, the author(s) used Chat GPT-4 in order to:Grammar and spelling check. Using this tool, the author(s) reviewed and edited the content as needed and take full responsibility for the publication’s control

References

[1] A. Hegde, G. Kavya, S. Coelho, H. L. Shashirekha, Mucs@ dravidianlangtech2023: leveraging learning models to identify abusive comments in code-mixed dravidian languages, in: Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages, 2023, pp. 266–274.

[2] S. Mishra, S. Mishra, 3idiots at hasoc 2019: Fine-tuning transformer neural networks for hate speech identification in indo-european languages., in: FIRE (working notes), 2019, pp. 208–213.

[3] A. Hegde, M. D. Anusha, S. Coelho, H. L. Shashirekha, B. R. Chakravarthi, Corpus creation for sentiment analysis in code-mixed Tulu text, in: M. Melero, S. Sakti, C. Soria (Eds.), Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages, European Language Resources Association, Marseille, France, 2022, pp. 33–40. URL: https://aclanthology.org/2022.sigul-1.5/.

[4] S. Coelho, A. Hegde, P. Lamani, H. L. Shashirekha, et al., Mucsd@ dravidianlangtech2023: Predicting sentiment in social media text using machine learning techniques, in: Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages, 2023, pp. 282–287.

[5] S. Sai, Y. Sharma, Towards offensive language identification for dravidian languages, in: Proceedings of the first workshop on speech and language technologies for Dravidian languages, 2021, pp. 18–27.

[6] A. Kumar, S. Saumya, A. Singh, Detecting dravidian offensive posts in miot: A hybrid deep learning framework, ACM Transactions on Asian and Low-Resource Language Information Processing (2023).

[7] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, et al., Dravidiancodemix: sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text, Language Resources and Evaluation 56 (2022) 765–806. URL: https://doi.org/10.1007/s10579-022-09583-7. doi:10.1007/s10579-022-09583-7.

[8] S. Divya, N. Sripriya, Transformer based model for offensive content recognition in dravidian languages, Brazilian Journal of Development 9 (2023) 30656–30667.

[9] N. Sripriya, B. R. Chakravarthi, T. Durairaj, B. Bharathi, S. C. Navaneethakrishnan, P. K. Kumaresan, A. M D, P. R. Hegde, D. Vikram, Overview of the shared task on offensive language identification in dravidian code-mixed languages, in: Forum of Information Retrieval and Evaluation FIRE-2025, 2025.