DravidianCodeMix 2025: Comparative Study of Transformer-based Models for Offensive Content Detection in Tamil, Malayalam, Kannada and Tulu Code-Mixed Texts

Santhiya P[1], Akshitha A V[1], Arul Murugan S[1] and Chandran T[1]

[1]*Kongu Engineering College, Tamil Nadu, India*

Abstract

An essential task in ensuring safe digital interactions, especially on social media platforms where multilingual exchanges are common, is the detection of offensive language in Dravidian code-mixed text. This project focuses on building classification models for Tamil, Malayalam, Kannada, and Tulu, four major Dravidian languages. To address this, we explored transformer-based approaches and evaluated their effectiveness across the datasets. Our study shows that IndicBERTv2-m2m was more effective for Tulu and Malayalam, whereas TwHIN-BERT yielded better outcomes for Tamil and Kannada. These observations emphasize that model suitability varies by language, highlighting the necessity of adopting language-specific strategies for offensive content detection. Furthermore, our work provides a foundation for handling low-resource language scenarios and code-mixed text challenges. The datasets also facilitate research into cross-lingual and multilingual learning approaches. Ultimately, these efforts contribute to safer online environments and more inclusive digital communication.

1. Introduction

The rapid growth of social media has made code-mixed communication—where regional languages blend with English—increasingly common. Dravidian languages such as Tamil, Malayalam, Kannada, and Tulu often appear in these exchanges, mixing scripts and vocabularies. While expressive, this also creates avenues for offensive content that disrupts online discourse. Detecting such language is difficult due to limited annotated data, inconsistent spellings, and translation issues. Recent shared tasks on offensive language identification [1] highlight these challenges and demonstrate the potential of transformer-based models. In this study, we evaluate offensive language detection using the DravidianCodeMix corpus [2] for Tamil, Malayalam, and Kannada, along with a Tulu dataset [3]. Our results show that IndicBERTv2-m2m performs best for Tulu and Malayalam, while TwHIN-BERT achieves superior performance for Tamil and Kannada, underscoring the need for language-specific model selection in real-world moderation.

2. Literature Survey

Research on offensive language identification in Dravidian code-mixed text has grown rapidly, largely driven by shared tasks and dataset creation. The FIRE 2020 and 2021 shared tasks provided early evaluations on Tamil, Malayalam, and Kannada, highlighting the challenges of detecting offensive expressions in code-mixed settings and motivating the use of both classical machine learning and multilingual transformer methods [4]. Later FIRE tasks expanded coverage to more languages and refined annotation schemes, enabling broader evaluation [1].

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

*Corresponding author.

santhiyaashok172@gmail.com (S. P); akshithaav.22cse@kongu.edu (A. A. V); arulmurugans.22cse@kongu.edu (A. M. S); chandrant.22cse@kongu.edu (C. T)

0000-0043-2073-426X (S. P); 0000-0123-27744-246X (A. A. V); 0000-0123-27744-246X (A. M. S); 0000-0123-27744-246X (C. T)

A key resource is the DravidianCodeMix dataset, which offers annotated corpora for Tamil, Malayalam, and Kannada [2]. It includes multiple offensive categories and reflects natural variation in social media text, such as Romanization and spelling inconsistencies, making it a realistic benchmark. More recently, a low-resource corpus for Tulu extended research to another Dravidian language, addressing severe data scarcity and enabling cross-lingual transfer studies [3]. Together, these datasets provide a foundation for multilingual modeling and sociolinguistic analysis of abusive discourse.

Beyond resources, model development has advanced performance. While early multilingual transformers achieved strong baselines, results varied by language due to dataset size and code-mixing complexity. To improve this, Chakravarthi et al. [5] introduced a multilingual MPNet and CNN fusion model for Tamil, Malayalam, and Kannada. Their hybrid architecture handled code-mixing effectively and outperformed both traditional machine learning and single-model transformers.

Overall, shared tasks and datasets such as DravidianCodeMix and Tulu have standardized evaluation and expanded coverage, while hybrid deep learning models have established robust baselines[5]. These developments provide a pathway for improving low-resource adaptation and advancing offensive language detection in multilingual social media.

3. Materials and Methods

3.1. Taskset Description

This work analyzes code-mixed datasets for four Dravidian languages—Tamil, Malayalam, Kannada, and Tulu—collected from social media platforms where English is often blended with regional languages. Each dataset is annotated as offensive or non-offensive, with subcategories distinguishing insults targeted at individuals, groups, or untargeted abuse. The Tamil dataset contains categories such as Not Offensive, Insult Individual, Insult Group, and Untargeted, while the Malayalam dataset adds an Other Language class for cross-lingual entries. Kannada follows a similar structure but introduces a Not-Kannada class to capture comments written in English, Hindi, or Romanized forms. The smaller Tulu dataset uses Not Offensive, Targeted, and Untargeted categories, providing insights into low-resource Tulu-English interactions. Related initiatives include the DOSA dataset for offensive span identification in Dravidian languages [6] and a recent corpus for Tulu offensive language detection [3], both underscoring the need to address low-resource challenges. These datasets not only enable systematic evaluation of offensive language detection methods but also reflect real-world issues like code-switching, inconsistent spellings, and dialectal variation. Collectively, they form a solid basis for advancing abusive content detection in Dravidian code-mixed contexts. Additionally, they support experimentation with multilingual and transformer-based models to handle mixed-language inputs effectively. They also provide opportunities to analyze sociolinguistic patterns and user behavior in online interactions. Future research can leverage these resources to improve cross-lingual transfer and low-resource learning strategies.

Table 1
Sample training texts from Tamil dataset

Text	Labels
14.12.2018 epo trailer pathutu irken ... Semaya iruku	Not_offensive
Paka thano poro movie la Enna irukunu	Not_offensive
“U kena tunggu lebih lama lagi untuk tahu saya” – chiyaan recognized	not-Tamil
Suriya anna vera level anna mass	Not_offensive
suma katththaatha da sound over a pooda kudaathu pa s3 1 month oda stop aakidum then bairavaa da aadchi than katthti katthti thondaiya kilikatha pa	Offensive_Untargetede

Table 2
Sample training texts from Malayalam dataset

Text	Labels
Muriyandikal pannikuutattam pole thurimezhukuvansllloosenes are rememmbering pulimurugan	Offensive_Targeted_Insult_Group
Nth bhashayado ith vech ketti malabr bhasha samsarikkunnu. Mohanlal malabarine kaliyakkunnu	Offensive_Targeted_Insult_Individual
Mere naam ka kachra karo sab mil k	not-malayalam
50 pathu pravasam kandavar adi like	Not_offensive
Padam kandavarkku like aam Mass ka baap	Not_offensive

Table 3
Sample training texts from Kannada dataset

Text	Labels
Yes yes yes	not-Kannada
Shankara alpatt kappuvettu, jeevana nadisutiruva uttama gunamattada vyakti	Not_offensive
Dislike madidorella bevarsi TikTokers..thika urkondavre Burnol kodrooo	Offensive_Targeted_Insult_Group
Yava game bro edu	Not_offensive
Hand's up	Not_offensive

Table 4
Sample training texts from Tulu dataset

Text	Labels
Nayi sulemagne Ee dialogue band South canara dalli helu obba appange huttidre	offensive targeted
enchi savuda film ye daravahi rd la kade	not offensive
Sonu nigam da Shreya ghoshal background g panka..	offensive untargeted
Tulu baarandinaaye judge dada malpve	not offensive
Dakshina Kannada anta ittirrodu enu shoki ga matte.	not tulu

3.2. Data Collection

The Data Collection module plays a key role in compiling code-mixed text in Tamil, Malayalam, Kannada, and Tulu from varied sources such as social media platforms, discussion forums, and publicly available repositories. This step is designed to capture a broad spectrum of linguistic features, including regional dialects, colloquial slang, and examples of both offensive and non-offensive language. Creating a diverse and balanced dataset is crucial, as it forms the foundation for training machine learning models capable of reliably identifying harmful content within multilingual and code-mixed contexts. In addition, proper sampling ensures fair representation of all classes, reducing bias in model predictions. The quality of data collected at this stage directly influences the accuracy and generalizability of the final system.

3.3. Data Preprocessing and Feature Extraction

Before being processed by machine learning models, the text must be standardized and transformed into a suitable format. The preprocessing stage cleans raw code-mixed text by removing noise such as URLs, special characters, emoticons, and extra whitespace, while also performing language-specific tasks like tokenization and stemming for both Dravidian and English components. This improves data consistency and prepares it for feature extraction.

After preprocessing, the text is converted into dense embeddings using transformer-based models such as IndicBERTv2-m2m and TwHIN-BERT. Unlike traditional approaches like TF-IDF or n-grams, transformers capture semantic and syntactic patterns across multiple languages and scripts, which is

crucial in code-mixed contexts where context often shifts between English and Dravidian. Leveraging these pretrained multilingual embeddings provides rich representations that strengthen the model's ability to distinguish offensive from non-offensive content.

3.4. Model Training and Selection

The Model Training and Selection phase focuses on fine-tuning transformer-based models with code-mixed datasets. In this work, IndicBERTv2-m2m is utilized for Tulu and Malayalam, while TwHIN-BERT is applied to Tamil and Kannada. Each model is trained with language-specific annotated corpora so that it can adapt to the unique traits of the respective code-mixed languages. The evaluation of model performance is carried out using common metrics such as accuracy, precision, recall, and F1-score. Based on these results, the most suitable transformer model is chosen for each language, enabling reliable identification of offensive and abusive expressions in Dravidian social media content. This process also helps reveal cross-lingual differences, showing how certain architectures generalize better to low-resource contexts. Moreover, careful model selection ensures that downstream applications, such as moderation systems, remain both efficient and scalable.

4. Results and Discussion

The experimental analysis of the proposed system indicates that abusive language detection yields varied outcomes across Dravidian languages, highlighting the importance of adopting models tailored to each language. For Tulu [3] and Malayalam, IndicBERTv2-m2m produced stronger results, whereas TwHIN-BERT achieved the highest accuracy on Tamil and Kannada text [7]. These differences arise from the unique linguistic structures, levels of code-mixing, and dataset characteristics associated with each language, demonstrating that no single algorithm consistently delivers the best performance across all cases. The results further reveal that thorough preprocessing combined with effective feature extraction substantially enhances model accuracy, enabling reliable identification of offensive content within multilingual and code-mixed social media. In addition, the findings emphasize the role of dataset balance, as skewed distributions tend to reduce performance on minority classes. These insights provide valuable guidance for building future models that can handle both linguistic diversity and resource scarcity more effectively.

4.1. Performance Analysis

The evaluation highlights persistent challenges in detecting offensive content across Dravidian code-mixed languages, largely due to the linguistic variability present in social media communication. Slang, acronyms, emojis, and region-specific informal expressions frequently disrupt semantic clarity, leading to misclassifications. Contextual ambiguity further complicates detection, as meanings shift based on discourse, speaker intent, or cultural nuance, causing both false positives and false negatives. Malayalam and Tulu suffer from small annotated datasets, limiting generalization and reducing robustness on unseen inputs. Tamil exhibits heavy code-switching between English and native scripts, producing multiple orthographic variants for the same term and making contextual understanding more difficult. Kannada faces additional issues such as spelling inconsistencies, mixed-script usage, and borrowing from neighboring languages, all of which introduce noise into classification.

These findings underscore the need for larger and more representative datasets, along with advanced context-aware architectures capable of modeling fine-grained linguistic cues. Incorporating linguistic tools such as morphological analyzers, character-level models, or subword embeddings may improve the handling of complex structures like agglutination, dialectal variations, and transliteration. Approaches such as cross-lingual transfer learning, data augmentation, and pretraining on region-specific corpora can further enhance robustness. Despite the limitations, the proposed system establishes a strong foundation for future research and offers practical insights for building effective moderation tools to manage offensive content in multilingual online platforms.

Table 5
Epoch-wise Performance of TwHIN-BERT on Tamil Dataset

Language	Model	Epoch	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Tamil	TwHIN-BERT	1	0.68	0.66	0.64	0.65
Tamil	TwHIN-BERT	2	0.70	0.68	0.67	0.67
Tamil	TwHIN-BERT	3	0.72	0.70	0.69	0.69

Table 6
Epoch-wise Performance of IndicBERTv2-m2m on Malayalam Dataset

Language	Model	Epoch	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Malayalam	IndicBERTv2	1	0.74	0.72	0.70	0.71
Malayalam	IndicBERTv2	2	0.77	0.75	0.74	0.74
Malayalam	IndicBERTv2	3	0.79	0.77	0.76	0.76

Table 7
Epoch-wise Performance of TwHIN-BERT on Kannada Dataset

Language	Model	Epoch	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Kannada	TwHIN-BERT	1	0.68	0.66	0.65	0.65
Kannada	TwHIN-BERT	2	0.70	0.68	0.67	0.67
Kannada	TwHIN-BERT	3	0.72	0.70	0.69	0.69

Table 8
Epoch-wise Performance of IndicBERTv2-m2m on Tulu Dataset

Language	Model	Epoch	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Tulu	IndicBERTv2	1	0.78	0.75	0.72	0.73
Tulu	IndicBERTv2	2	0.81	0.79	0.77	0.78
Tulu	IndicBERTv2	3	0.83	0.81	0.80	0.80

4.2. Error Analysis

The performance evaluation highlights several challenges in detecting objectionable content across code-mixed Dravidian languages. Slang, acronyms, and region-specific informal expressions often cause misclassifications, while contextual ambiguity leads to false positives and negatives when words shift meaning across situations. For Malayalam and Tulu, the limited size of annotated datasets reduces generalization and weakens performance on unseen data. In Tamil, frequent code-switching between English and native scripts complicates context detection, while Kannada suffers from spelling inconsistencies, mixed-script usage, and borrowed terms from neighboring languages, all of which add noise. These issues emphasize the need for larger, more representative datasets and advanced context-aware models to improve detection accuracy. Nevertheless, the system provides a strong starting point for future research and practical control of offensive content in multilingual online platforms.

5. Limitations

The proposed offensive language detection method performs well but still faces limitations. It depends on small annotated datasets that fail to capture regional dialects, slang, and code-mixed expressions in Tamil, Malayalam, and Tulu. Ambiguous sentences can lead to misclassification, and rapidly changing, informal social media language may further reduce accuracy. Larger, more diverse datasets and advanced context-aware methods are needed to improve robustness and generalization.

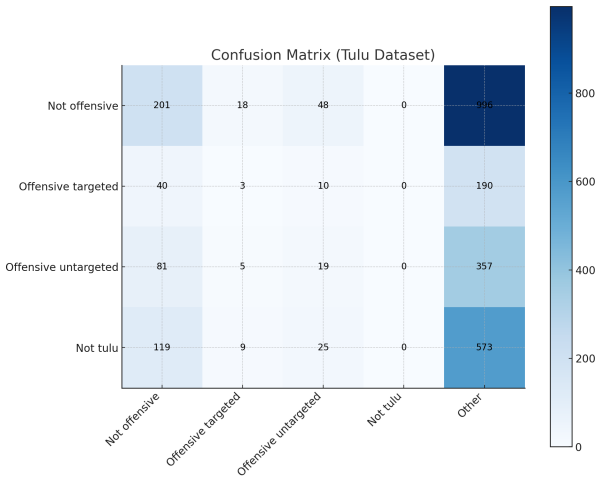

Figure 1: Confusion matrix for the Tulu dataset

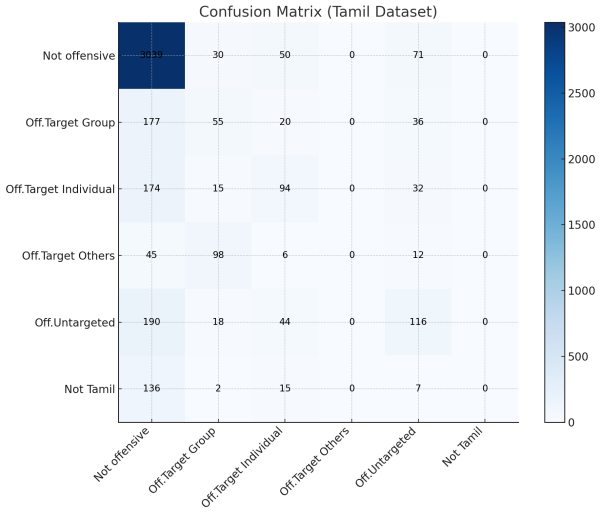

Figure 2: Confusion matrix for the Tamil dataset

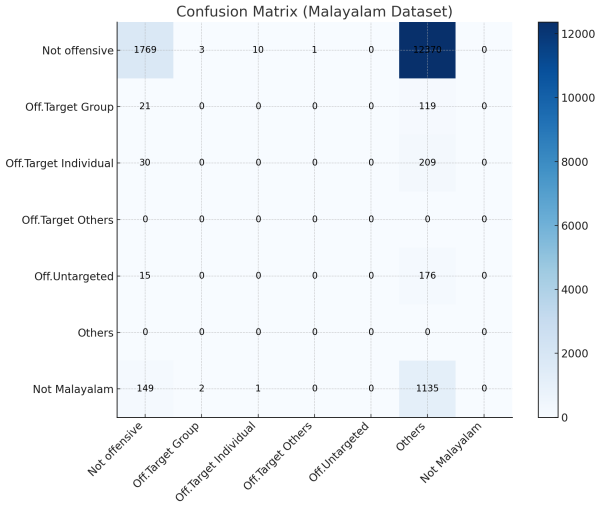

Figure 3: Confusion matrix for the Malayalam dataset

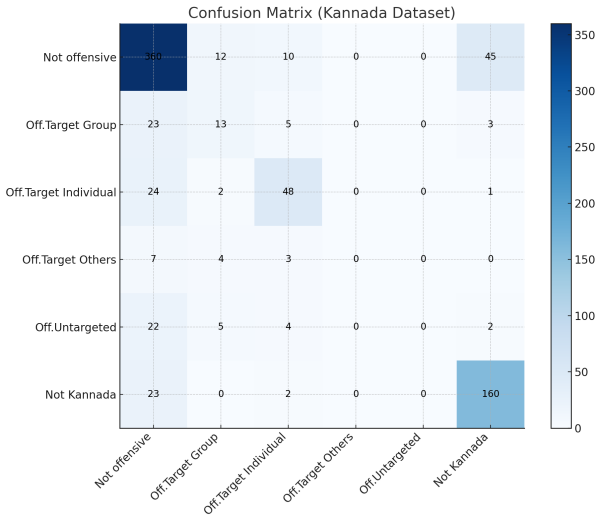

Figure 4: Confusion matrix for the Kannada dataset

6. Conclusion

Evaluation outcomes demonstrate that the performance of abusive language detection varies across Dravidian languages, emphasizing the importance of selecting models tailored to individual linguistic contexts [1]. IndicBERTv2-m2m achieved higher accuracy for Tulu and Malayalam [3], whereas TwHIN-BERT performed better for Tamil and Kannada [7]. The suggested system offers a solid basis for automatic content filtering, even though some mistakes still occur because of unclear context, slang, and small datasets. All things considered, this work promotes safer online communication and establishes the framework for next advancements, such as context-aware algorithms and larger datasets for more reliable foul language identification [2].

Project Repository

The full source code for this project is available on
GitHub: GitHub Repository- Chandrant-chan

Declaration on Generative AI

In the course of preparing this manuscript, the author(s) employed the generative AI tool ChatGPT. Its use was limited to performing checks for grammar and spelling. Following this, the author(s) conducted a thorough review and revision of the text and assume full responsibility for the final published content.

References

[1] N. Sripriya, B. R. Chakravarthi, T. Durairaj, B. Bharathi, C. N. Subalalitha, P. K. Kumaresan, M. D. Anusha, 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.

[2] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, N. Jose, S. Suryawanshi, E. Sherly, J. P. McCrae, Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text, Language Resources and Evaluation 56 (2022) 765–806.

[3] A. M. D, D. Vikram, B. R. Chakravarthi, P. R. Hegde, Overcoming low-resource barriers in tulu: Neural models and corpus creation for offensive language identification, 2025. URL: https://arxiv.org/abs/2508.11166. arXiv:arXiv:2508.11166.

[4] B. R. Chakravarthi, R. Priyadharshini, N. Jose, T. Mandl, P. K. Kumaresan, R. Ponnusamy, J. P. McCrae, E. Sherly, et al., 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, 2021, pp. 133–145.

[5] B. R. Chakravarthi, M. B. Jagadeeshan, V. Palanikumar, R. Priyadharshini, Offensive language identification in dravidian languages using mpnet and cnn, International Journal of Information Management Data Insights 3 (2023) 100151. URL: https://www.sciencedirect.com/science/article/pii/S2667096822000945. doi:https://doi.org/10.1016/j.jjimei.2022.100151.

[6] B. R. Chakravarthi, et al., Dosa: Dravidian code-mixed offensive span identification dataset, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, 2021.

[7] S. Roy, et al., Hottest: Hate and offensive content identification in tamil using deep learning, in: Proceedings of DravidianLangTech Workshop, 2023.