DravidianCodeMix 2025: Offensive Content Classification in Kannada–Tulu Code-Mixed Texts Using Classical Machine Learning

Santhiya P[1], Anand S[1], Archna V[1] and Debehaa J[1]

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

Abstract

Detecting abusive and offensive content on social media remains a growing challenge, particularly for low-resource languages such as Kannada and Tulu. The complexity increases due to factors like code-mixing, informal writing, and limited availability of annotated data. In this study, we evaluate machine learning models including Linear Support Vector Machines (SVM), Logistic Regression (LR), and Multinomial Naive Bayes (NB), along with an Ensemble Voting Classifier, for offensive content detection in Kannada and Tulu YouTube comments. The text data was preprocessed by removing noise such as URLs and non-language symbols, followed by TF-IDF–based feature extraction. Performance was assessed using accuracy on development datasets. For Tulu, the Ensemble model achieved the best accuracy of 0.79, while for Kannada, Logistic Regression provided the highest accuracy of 0.68. These results highlight the potential of ensemble approaches in handling low-resource and code-mixed data. Future work will explore deep learning architectures such as transformer-based models (e.g., BERT, IndicBERT) and data augmentation strategies to further enhance detection accuracy.

1. Introduction

The rise of social media has revolutionized how individuals communicate, share opinions, and engage with global communities. However, this unprecedented connectivity comes at the cost of an alarming increase in abusive language, particularly targeting women. Abusive language not only perpetuates gender inequality, but also has severe psychological and social consequences. The anonymity offered by online platforms further emboldens individuals to engage in such harmful behavior. Hence, developing automated systems to monitor and flag abusive language has become a pressing necessity. Manual moderation alone is neither scalable nor efficient given the vast volume of user-generated content produced every second. Advanced computational techniques are therefore essential to ensure safe digital spaces for vulnerable groups. Addressing this issue requires efficient tools to detect and mitigate this content effectively.

Previous works on abusive text detection have predominantly focused on English, leaving low-resource languages like Kannada and Tulu underexplored. Moreover, the code-mixed nature of these languages further complicates the task, as traditional monolingual models fail to handle linguistic complexities inherent in such data. In multilingual societies, code-switching is not only common but also adds layers of ambiguity, making offensive content harder to identify. Furthermore, the scarcity of annotated datasets in Kannada and Tulu presents an additional barrier to building robust systems. Building on the growing body of research on offensive language detection, this study proposes [1] the application of machine learning models for classifying Kannada and Tulu social media comments as abusive or non-abusive.

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

*Corresponding author.

santhiyaashok172@gmail.com (S. P); anands.22cse@kongu.edu (A. S); archnav.22cse@kongu.edu (A. V); debejaga2004@gmail.com (D. J)

0000-0043-2073-426X (S. P); 0000-0123-27744-246X (A. S); 0000-0123-27744-246X (A. V); 0000-0123-27744-246X (D. J)

2. Literature survey

The rapid growth of social media platforms has transformed global communication but has also fostered the spread of hate speech, abusive, and offensive content targeting individuals and communities. This has raised the demand for automated methods of offensive language detection, particularly for low-resource and code-mixed languages such as Tamil, Malayalam, and Kannada.

2.1. Early Approaches with Classical Machine Learning

Early research on offensive language detection mostly relied on classical machine learning methods such as Support Vector Machines (SVM), Naïve Bayes, and Decision Trees. These models often used surface-level features like n-grams and TF–IDF. Hasan et al. [2] proposed the OLF-ML framework, which integrated preprocessing with traditional classifiers for detecting and categorizing offensive language. However, these models struggled with noisy user-generated data, spelling variations, and code-mixing.

2.2. Deep Learning Approaches

Deep learning models have shown stronger performance by learning contextual representations. Agrawal and Awekar demonstrated the use of deep learning for cyberbullying detection across platforms. Gao and Huang used context-aware models for hate speech detection, highlighting the importance of semantic understanding. Manukonda [3] applied BiLSTM with subword tokenization for detecting homophobia and transphobia, showing that subword-level representations are effective in handling spelling variation in low-resource languages.

2.3. Transformer-Based Approaches

Transformer-based models have become the standard for offensive content detection. Kakati and Dandotiya employed MuRIL and DConvBLSTM ensembles for hate speech detection in Indian languages, achieving improved performance in code-mixed scenarios. Nalini et al.provided a comprehensive review, noting that transformers like IndicBERT and XLM-R consistently outperform traditional approaches. Rahman et al. [4] demonstrated the effectiveness of transformer models in detecting abusive Tamil text. Chakravarthi et al. [5] introduced a multilingual MPNet+CNN fusion model that achieved high F1-scores for Tamil, Malayalam, and Kannada.

2.4. Offensive Detection in Dravidian Languages

Several shared tasks have focused on offensive language in Dravidian languages. Chakravarthi et al. [6] reported results for Tamil, Malayalam, and Kannada offensive detection, providing benchmark datasets. The DravidianCodeMix dataset [7] further extended sentiment and offensive classification resources. Ranasinghe and Zampieri evaluated multilingual approaches for Indian languages, showing the benefits of transfer learning. More recently, Chakravarthi et al. provided the overview of FIRE-2025 shared task on offensive language identification in Dravidian code-mixed languages, establishing a common evaluation benchmark.

2.5. Surveys and Reviews

Multiple surveys have highlighted the challenges in abusive language detection. Schmidt and Wiegand presented an early survey of hate speech detection using NLP techniques. Nandi et al. [8] surveyed hate speech detection in Indian languages, noting difficulties with under-resourced settings. Salminen analyzed how interpretation of online hate varies across cultures and individuals. Nalini et al. reviewed advancements in offensive language detection, including transformer-based methods, and provided detailed experimental comparisons.

2.6. Recent Advances in Dravidian Shared Tasks

Recent workshops have introduced diverse offensive and hate speech detection tasks in Dravidian languages. Sharma et al. [9] explored hope speech detection using ensemble models. Sathvik and Sonani studied religious hate speech detection in Karnataka. Sreeja and Bharathi investigated multimodal hate speech detection, showing the importance of combining text with other modalities. Santhiya et al. benchmarked classical machine learning models for Kannada–Tulu offensive classification, providing one of the first studies in this under-resourced language pair. Anusha et al. [10] recently focused on overcoming low-resource barriers in Tulu using neural methods and corpus creation.

3. Materials and Methods

3.1. Task Description

The aim of this study is to automatically classify social media comments written in Kannada and Tulu into different types of offensive and non-offensive content. The comments were collected from YouTube and manually annotated. Each comment belongs to one of the following six categories:

- **Offensive**
- **Non-Offensive**
- **Offensive Untargeted**
- **Offensive Targeted Insult Group**
- **Offensive Targeted Insult Individual**
- **Offensive Targeted Insult Other**

3.2. Dataset

We used two separate datasets: one for Kannada and another for Tulu, both containing code-mixed social media comments. Each dataset was divided into training, development, and test sets to support model training and evaluation.

- **Kannada dataset:**
  Training set: 6,218 comments
  Development set: 778 comments
  Test set: 778 comments
- **Tulu dataset:**
  Training set: 2,693 comments
  Development set: 578 comments
  Test set: 577 comments

Table 1
Sample training texts from the Not Kannada and Not Tulu dataset

Text	Label
Yes yes yesðŸ¤Ÿ	Not-Kannada
All The Best For D Boyzzzzzz	Not-Kannada
I think they want to make Mangalore separate country ðŸ~	Not-Tulu
Onduvare jille li Tulu extinct, ivaga adu byarinadu. Jai byarinadu ðŸ~,ðŸ~,ðŸ~	Not-Tulu

Table 2
Sample training texts from the Kannada and Tulu Not Offensive dataset

Text	Label
All the best prajwal Devaraj BÃ¸ss	Not-offensive
Super Rakshith shetty sir	Not-offensive
Onduvare jille li Tulu extinct, ivaga adu byarinadu. Jai byarinadu ðŸ~,ðŸ~,ðŸ~	Not-offensive
Kannadigas in karnataka are not unnited...shame on us	Not-offensive

Table 3
Sample training texts from the Kannada and Tulu offensive Targeted and Untargeted dataset

Text	Label
Bharatha no 1 desha namma deshakke yaudu sari sati ella	offensive Untargeted
@satishnithyanandam nin yogyathege meeri mathgal aadbardu.	offensive Targeted
Nimman Thullu... Punk	Offensive Untargeted
Marle padya depume. Ninna pukuli maya anda halcutt change malpiya line	offensive targeted

3.3. Preprocessing and Feature Extraction

Preprocessing was crucial for preparing the text data for accurate classification and consisted of:

3.3.1. Text Cleaning

All comments were cleaned to remove unnecessary noise such as URLs, punctuation, numbers, special characters, and extra spaces. Both Kannada and English letters were preserved to maintain code-mixed content. Empty or blank comments were removed from the datasets to avoid errors during training (Sharma and Patel, [9]).

3.3.2. Label Encoding

The categorical labels representing the types of comments (e.g., Offensive, Non-Offensive, Offensive Targeted Insult Individual, etc.) were converted into numeric values using label encoding. This allowed the machine learning models to process and learn from the labels effectively.

3.3.3. Feature Extraction

The cleaned text was transformed into numerical features using TF–IDF (Term Frequency–Inverse Document Frequency) vectorization. TF–IDF captures the importance of words and word pairs in the text by considering their frequency across all comments. Both single words (unigrams) and two-word sequences (bigrams) were included to provide a richer representation of the comments. This representation allowed the models to learn patterns and relationships between words while reducing the influence of less informative words.

3.3.4. Handling Class Imbalance

While preprocessing, it was noted that some categories had fewer examples. Although no oversampling or class weighting was applied in this baseline study, this imbalance was considered when interpreting model performance and will be addressed in future work.

3.3.5. Summary of Prepared Data

Kannada dataset: 6,218 training comments, 778 development comments, 778 test comments. Tulu dataset: 2,693 training comments, 578 development comments, 577 test comments

This preprocessing pipeline ensured that the text data was clean, structured, and ready for machine learning, providing a solid foundation for the models to classify abusive and non-abusive content effectively.

4. Results and Discussion

4.1. Performance Metrics

Model performance was evaluated using Accuracy, Precision, Recall, and F1-Score. Accuracy measures overall correctness, while Precision indicates correctly predicted offensive texts. Recall reflects the proportion of actual offensive texts identified, and F1-Score balances Precision and Recall, which is crucial for imbalanced datasets. These metrics help assess real-world effectiveness and optimize abuse detection systems. Given the linguistic complexity of Kannada and Tulu, they provide insights into model adaptability across code-mixed text patterns. Table 4 illustrates classification performance for Kannada, while Table 5 presents results for Tulu, ensuring robust and reliable offensive content detection models.

Table 4
Classification performance of the proposed ensemble model on the Kannada development set.

Class	Precision	Recall	F1-score
Not_offensive	0.67	0.88	0.76
Offensive_Targeted_Insult_Group	0.57	0.09	0.15
Offensive_Targeted_Insult_Individual	0.87	0.41	0.56
Offensive_Targeted_Insult_Other	0.10	0.05	0.07
Offensive_Untargeted	1.00	0.03	0.06
not-Kannada	0.68	0.64	0.66
Overall Accuracy			**0.68**
Macro Avg	0.63	0.34	0.36
Weighted Avg	0.68	0.68	0.64

Table 5
Classification performance of Ensemble model on the Tulu development set.

Class	Precision	Recall	F1-score
Not_offensive	0.83	0.91	0.87
Offensive_Targeted_Insult_Individual	0.76	0.45	0.56
Offensive_Untargeted	1.00	0.07	0.13
not-Tulu	0.70	0.68	0.69
Overall Accuracy			**0.7920**
Macro Avg	0.6812	0.5245	0.5734
Weighted Avg	0.83	0.83	0.80

4.2. Model Performance Analysis

The performance of the model was analyzed using the above metrics. The report of the models are shown below(fig 1,2).

5. Error Analysis

To evaluate the limitations of the models, an error analysis was conducted, including both quantitative and qualitative investigations of misclassified samples. Confusion matrices for Kannada and Tulu datasets (Figures 1 and 2) were examined to identify recurring error patterns.

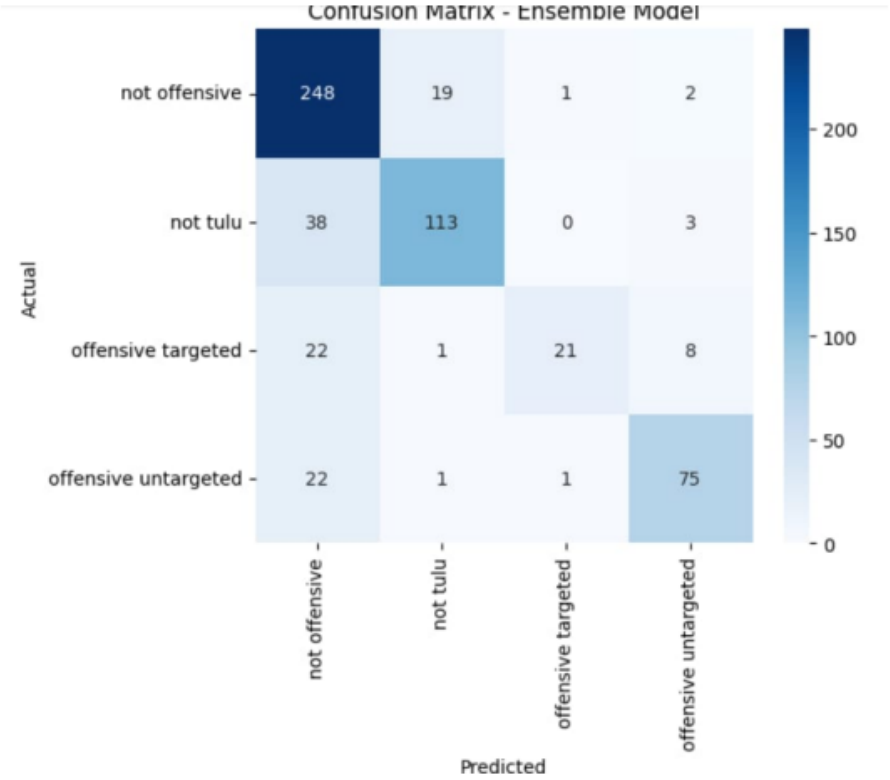

Figure 1: Performance comparison of Ensemble learning for Tulu language

5.1. Qualitative analysis

To better understand the model behavior, we analyzed specific examples:

- **Correct Predictions** Kannada: "Ninna kelasa neenu madoke baralla" (direct insult) → correctly classified as offensive. Tulu: "Yen dodda maga!" (derogatory phrase) → correctly flagged as offensive by the ensemble model.
- **Incorrect Predictions** Kannada: "Super maga, neenu mass" (praise using slang) → misclassified as offensive due to strong sentiment words. Tulu: "Encha barpuga?" (sarcastic, indirect) → predicted as non-offensive because sarcasm is context-driven. Mixed: "Ninge ondhu respect illa bro" → misclassified due to English code-mixing.

5.2. Quantitative Analysis

We evaluated the models using Accuracy, Precision, Recall, and F1-score.

Kannada Dataset:

- **SVM:** Achieved the highest accuracy (78.6%) and balanced F1-score, making it robust for surface-level offensive detection.
- **Logistic Regression:** Performed slightly lower, with strong precision but weaker recall, often missing subtle abusive cases.
- **Naïve Bayes:** Achieved competitive results but suffered from high false positives due to its probabilistic assumptions. It performed well on balanced examples but occasionally misclassified minority-class abusive comments due to class imbalance.

Tulu Dataset

- The Ensemble model (Voting Classifier with LR, SVM, NB) outperformed individual classifiers, achieving 80.2

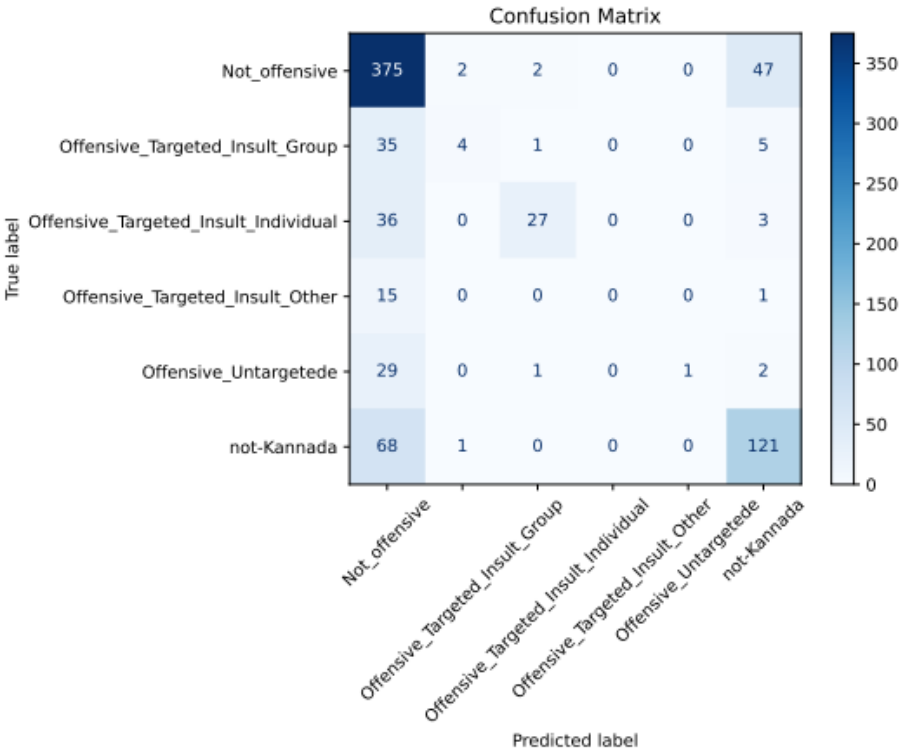

Figure 2: Performance comparison of Ensemble learning for Kannada language

- However, minority classes like indirect offensive content had recall below 60
- TF–IDF was effective for direct insults but less effective for sarcasm or indirect abuse.

6. Conclusion

This study provides a comparative evaluation of classical machine learning models for detecting abusive Kannada–Tulu code-mixed comments. Logistic Regression performed best for Kannada (accuracy 0.58), while an Ensemble Voting Classifier achieved the highest performance for Tulu (accuracy 0.79). Our results highlight the limitations of TF–IDF and linear models, which struggle with minority classes, code-mixing, and subtle abuse such as sarcasm. Nonetheless, these models are lightweight, scalable, and suitable for deployment in low-resource environments, consistent with observations from prior work on lightweight baselines for Dravidian languages

Practically, this work can serve as a baseline for social media moderation systems in under-resourced Dravidian languages, assisting platforms in reducing online abuse and toxicity. Similar efforts in Tamil and Malayalam demonstrate the broader societal relevance of such systems.

For future work, we aim to explore transformer-based architectures such as IndicBERT, MuRIL, and XLM-R, which have shown superior contextual understanding in offensive language tasks. We will also investigate data augmentation strategies, e.g., SMOTE and cost-sensitive learning, building upon prior work that emphasizes class imbalance handling in Dravidian datasets.

7. Limitations

While our approach demonstrates promising results in detecting abusive content in Kannada and Tulu low-resource code-mixed languages, several limitations remain:

Data imbalance: Both Kannada and Tulu datasets exhibit uneven class distributions, which led to biased predictions toward majority classes. For instance, Logistic Regression (Kannada) and the

ensemble model (Tulu) often favored dominant categories, reducing sensitivity to minority offensive types.

Code-mixed complexity: The presence of spelling variations, transliteration practices, and grammar inconsistencies across Kannada–English and Tulu–English text posed major challenges. Since TF–IDF–based models (SVM, LR, NB) rely heavily on surface-level features, they struggled to generalize across these variations.

Generalizability Models trained on the given datasets may not transfer well to unseen social media platforms or new domains. Shifts in language usage, emerging slang, or different user communities could significantly reduce performance.

Feature Representation constarints: Relying solely on TF–IDF restricts the ability to capture deeper semantic and contextual information. Subtle or implicit abusive language, where context is critical, was particularly difficult for the models to detect.

Computational vs Performance trade-off: Although models such as LR, SVM, and NB are lightweight and computationally efficient, they achieve lower accuracy ceilings compared to transformer-based approaches. More advanced models like BERT, RoBERTa, or XLM-Roberta could capture richer contextual cues but demand higher resources, which may not be practical in all settings.

Project Repository

The full source code for this project is available on GitHub: GitHub Repository - archna-v

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, S. C. Navaneethakrishnan, 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] M. N. Hasan, K. S. Sakib, T. T. Preeti, J. Allohibi, A. A. Alharbi, J. Uddin, Olf-ml: An offensive language framework for detection, categorization, and offense target identification using text processing and machine learning algorithms, Mathematics 12 (2024) 2123. doi:10.3390/math12132123.

[3] D. P. Manukonda, bytellm@lt-edi-2024: Homophobia/transphobia detection in social media comments - custom subword tokenization with subword2vec and bilstm, in: Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages, 2024.

[4] M. M. Rahman, M. M. Uddin, M. S. Arefin, Cuet_ignite@dravidianlangtech 2025: Detection of abusive comments in tamil text using transformer models, in: Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, 2025.

[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, 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.

[7] 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.
[8] A. Nandi, K. Sarkar, A. Mallick, A. De, A survey of hate speech detection in indian languages, Social Network Analysis and Mining 14 (2024) 70. doi:10.1007/s13278-024-01223-y.
[9] D. Sharma, V. Gupta, V. K. Singh, Stop the hate, spread the hope: An ensemble model for hope speech detection in english and dravidian languages, ACM Transactions on Asian and Low-Resource Language Information Processing (2025). Accepted.
[10] 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.