OffenSwitch: Decoding Toxicity in Dravidian Code-Mixing with Transformers

Krishna Tewari[1,*], Supriya Chanda[2] and K Abhinay Paul[1]

[1] *Indian Institute of Technology (BHU), Varanasi, INDIA*
[2] *Bennett University, Greater Noida, INDIA*

Abstract

Offensive language detection is a vital task in natural language processing, especially given the rise of multilingual code-mixed text on social media platforms. This paper presents a shared task on offensive language identification in Dravidian code-mixed languages: Tamil-English, Malayalam-English, Kannada-English, and Tulu-English organized as part of FIRE 2025. The task aims to address the unique challenges posed by code-switching, morphological richness, and the use of non-native scripts endemic to Dravidian languages. Participants are required to classify social media comments into categories including offensive, not-offensive, targeted offensive, and untargeted offensive (group / individual). We provide a gold-standard dataset curated from YouTube comments on diverse topics and encourage the development of models capable of robust offensive language detection across these low-resource, multilingual settings. Our team, IReL@IIT-BHU, participated using fine-tuned XLM-RoBERTa models with and without early stopping. Across the four languages, our systems consistently ranked within the top-6, achieving 4th place in Kannada and Tamil, 5th in Malayalam, and 6th in Tulu, thereby establishing multilingual transformers as a strong baseline for Dravidian code-mixed offensive language identification.

Keywords

Code-Mixed, Offensive, Hate, NLP, Dravidian, Tamil, Kannada, Tulu, Malayalam

1. Introduction

Offensive language detection has become a critical task in Natural Language Processing (NLP), particularly in the era of rapid digital communication where user-generated content proliferates on social media platforms. Online spaces such as YouTube, Facebook, and Twitter host millions of daily discussions that range from entertainment to politics and social issues. While these platforms foster open dialogue, they also provide fertile ground for the spread of offensive content, ranging from abusive language and hate speech to cyberbullying and targeted harassment. Left unchecked, such behavior can escalate, harming individuals, polarizing societies, and undermining the safety of online communities.

The growing concern over offensive language has motivated the development of automatic detection systems. However, this task is far from trivial due to several challenges. Offensive expressions can be subtle, context-dependent, and highly variable across communities. Furthermore, multilingual societies introduce an additional layer of complexity: *code-mixing*, a phenomenon in which speakers switch between two or more languages within the same sentence or discourse. Code-mixed data is widespread in multilingual regions such as South India, where languages like Tamil, Malayalam, Kannada, and Tulu are often mixed with English in informal communication. Detecting offensive language in such data is challenging for traditional NLP models, which are predominantly trained on monolingual corpora and fail to capture the dynamics of code-switching across linguistic levels (lexical, morphological, and syntactic). The issue is compounded when users write these languages in non-native scripts, further complicating text normalization and understanding.

To address these challenges, the FIRE 2025 shared task focuses on offensive language identification in code-mixed Dravidian languages: Tamil-English, Malayalam-English, Kannada-English, and Tulu-

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

[*] Corresponding author.

krishnatewari.rs.cse24@itbhu.ac.in (K. Tewari); suplife24@gmail.com (S. Chanda); kabhinay.paul.cse20@itbhu.ac.in (K. A. Paul)
0009-0005-6599-9956 (K. Tewari); 0000-0002-6344-8772 (S. Chanda)

English [1]. Participants are provided with a gold-standard dataset curated from YouTube comments spanning news, entertainment, and socio-political discussions. The task requires classifying each instance into categories such as offensive, not-offensive, targeted offensive, and untargeted offensive (group / individual). The intended applications are wide-ranging, including assisting social media platforms in content moderation, supporting law enforcement agencies in identifying online abuse, and enabling brands to monitor consumer sentiment while maintaining a civil discourse. Beyond immediate use cases, the task aims to foster the development of robust multilingual and code-mixed NLP models that can generalize effectively across languages and cultures.

The rest of this paper is organized as follows: Section 2 discusses related work; Section 3 describes the dataset; Section 4 presents our proposed methodology; Section 5 reports results and analysis; and Section 6 concludes with key findings and future directions.

2. Related Work

Initial efforts in offensive language detection focused on monolingual English texts, employing techniques such as machine learning on tweets and forum discussions to model hate and offensive content [2, 3]. As social media environments have become increasingly multilingual, researchers have recognized the unique linguistic challenges presented by *code-mixed* text. The HASOC shared task and subsequent FIRE challenges demonstrated that systems designed for monolingual data often underperform on Hindi-English and Bengali-English code-mixed datasets, underscoring the need for methods tailored to linguistic mixing and complex orthographic patterns [4, 5].

For Hindi-English code-mixing, Bohra et al. developed a dedicated dataset for hate speech detection and experimented with methods including CRFs, SVMs, and neural models, highlighting the necessity of code-mixed resources and hybrid features [6]. Mathur et al. extended this line of work by focusing on offensive tweet classification with both deep learning and attention-based approaches, observing that classical and neural architectures each capture different facets of Hinglish data [7]. Transformer-based and multilingual models such as BERT, mBERT, and XLM-R have since demonstrated state-of-the-art results in Hindi-English code-switching scenarios, with studies confirming the critical advantages of transfer learning and contextual embeddings [8].

Barman et al. studied language identification as a precursor for downstream classification, finding that code-mixed Bengali-English presents further segmentation and tokenization hurdles, which can be partially addressed through subword models and character-level representations [9]. More recently, Chakravarthi et al. investigated multilingual embeddings and transfer learning, finding consistent improvements in both sentiment analysis and hate speech detection in low-resource Indo-Aryan and Dravidian code-mixed settings [10].

Within the Dravidian language context, Chakravarthi et al. developed new benchmarks for Tamil-English, Malayalam-English, and Kannada-English code-mixed data, as part of initiatives such as DravidianCodeMix and FIRE shared tasks [11, 12, 13]. These works established protocols for offensive content annotation and showcased the potential of deep learning, including CNNs and MPNet, for handling morphological and script variation in these under-resourced languages [14].

Chanda and Pal's early system at FIRE 2020 addressed sentiment analysis for code-mixed social media text using both classical and neural models, identifying data sparsity and script diversity as core bottlenecks [15]. Building upon this, Chanda et al. demonstrated that fine-tuning pre-trained transformer models yielded superior results for hate speech detection in Indo-Aryan and code-mixed contexts, motivating wider adoption of multilingual architectures [16]. Their later work leveraged multilingual embeddings for fine-grained conversational hate speech detection, successfully distinguishing nuanced and context-dependent offenses in Dravidian social streams [17]. The group also tackled related tasks such as sarcasm detection in code-mixed Tamil-English and Malayalam-English, and highlighted preprocessing innovations like effective stopword removal [18, 19, 20]. Furthermore, their deep learning-based approaches for hate speech detection in Tulu-English and other low-resource Dravidian pairs expanded both the data and methodological frontiers for safer digital spaces [21, 22].

The FIRE 2025 shared task extends these advances by releasing curated, gold-standard datasets for

Dravidian code-mixed offensive language detection and by promoting benchmark evaluation of neural models for Tulu, Tamil, Malayalam, and Kannada paired with English [23, 24]. These benchmark efforts include advanced methods based on MPNet and CNNs [14], and motivate continued research in robust code-mixed NLP for South Asian languages.

3. Dataset

The organizers of the shared task on Offensive Language Identification in Dravidian Code-Mixed Languages at FIRE 2025 provided annotated datasets for four language pairs: Tulu-English, Kannada-English, Malayalam-English, and Tamil-English. For each language, the data was divided into training, development, and testing splits. Each instance in the dataset consists of a code-mixed social media sentence and a corresponding offensive language label.

For the Tulu-English dataset, four categories were defined. The detailed statistics for this dataset are reported in Table 1. It can be observed that the class distribution is imbalanced, with the *not_offensive* category being the majority class.

For the other three language pairs Kannada-English, Malayalam-English, and Tamil-English the label set was more fine-grained. Multiple types of offensive expressions were included. The statistics of these datasets are summarized in Table 2. Here too, the data is skewed towards the *Not_offensive* category, especially in the case of Malayalam and Tamil.

The datasets reflect the natural distribution of code-mixed conversations on social media, where non-offensive instances dominate but offensive targeted insults are of particular interest for downstream classification. This imbalance motivated us to adopt transformer-based methods that are robust to such skewed label distributions.

Table 1
Dataset statistics for Tulu-English across training, development, and testing splits.

Label	Train	Dev	Test
not_offensive	1261	270	270
not_tulu	726	156	155
offensive_untargeted	462	99	99
offensive_targeted	243	52	52

Table 2
Dataset statistics for Kannada-English, Malayalam-English, and Tamil-English across training, development, and testing splits. Here, *not-{Language}* indicates non-code-mixed instances, where {Language} could be Kannada, Malayalam, or Tamil.

Label	**Kannada-English**			**Malayalam-English**			**Tamil-English**		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
Not_offensive	3544	426	427	14153	1779	1765	25425	3193	3190
not-{Language}	1522	191	185	1287	163	157	1454	172	160
Offensive_Untargeted	212	33	33	191	20	29	2906	356	368
Offensive_Targeted_Insult_Individual	487	66	75	239	24	27	2343	307	315
Offensive_Targeted_Insult_Group	329	45	44	140	13	23	2557	295	288
Offensive_Targeted_Insult_Other	123	16	14	–	–	–	454	65	71

4. Methodology

In this work, we focus on the task of offensive language identification in Tulu-English code-mixed data as part of the FIRE 2025 shared task. The objective is to classify each sentence into one of four categories: *offensive*, *not-offensive*, *targeted offensive*, or *untargeted offensive (group / individual).* To address this, we adopt a transformer-based architecture leveraging the multilingual capabilities of XLM-RoBERTa.

4.1. Data Preprocessing

The dataset provided by the shared task organizers was divided into training, validation, and test splits. Each sentence was paired with a corresponding label indicating its offensive language category. We employed the `LabelEncoder` from the scikit-learn library to map the textual labels into numerical representations. The text data was then tokenized using the `XLMRobertaTokenizer`, with a maximum sequence length of 128 tokens. Tokenization produced `input_ids` and `attention_mask`, which were padded and truncated as necessary.

4.2. Model Architecture

We fine-tuned the `xlm-roberta-base` model released by Facebook AI Research. The model was extended with a classification head to predict across the four output labels. We implemented a PyTorch-based dataset wrapper to efficiently handle tokenized sequences and labels during training and evaluation.

4.3. Training Strategy

The model was fine-tuned using the AdamW optimizer with a learning rate of 1×10^{-6} and a batch size of 32. The loss function used was cross-entropy loss, suitable for multi-class classification. We trained the model for up to 100 epochs while applying early stopping with a patience of 10 epochs to prevent overfitting. At each epoch, training and validation losses, along with accuracies, were recorded. The best performing model checkpoint on the validation set was retained for final evaluation.

4.4. Evaluation and Prediction

The trained model was evaluated on the held-out test set, and standard classification metrics such as precision, recall, and F1-score were reported using the scikit-learn `classification_report` function. Additionally, the fine-tuned model was used to generate predictions on the unlabeled data provided in the shared task. The model outputs were mapped back to the original label names using the fitted label encoder. We also plotted the training and validation curves for loss and accuracy across epochs to visualize the learning behavior of the model.

Overall, this methodology leverages the multilingual contextual representations of XLM-RoBERTa, combined with careful preprocessing and early stopping, to robustly address the challenge of offensive language identification in code-mixed Tulu-English social media text.

5. Results

The performance of our submitted systems was evaluated on four Dravidian code-mixed languages, namely Tulu, Kannada, Malayalam, and Tamil. The organizers provided the official leaderboard results, as shown in Tables 3-6. Our team participated under the name **IReL@IIT-BHU** with two submitted runs: Run 1 (without early stopping) and Run 2 (with early stopping).

In the Tulu offensive language identification task, our system achieved an mF1 score of 0.710, securing the **6th rank** overall (Table 3). We observed that applying early stopping helped to stabilize validation performance during training, but the improvement in the final test score was not substantial compared to the baseline run.

For the Kannada dataset, our submission with early stopping (Run 2) obtained an mF1 score of 0.430, placing us at the **4th rank** (Table 4). This indicates that our approach generalized reasonably well to Kannada, with performance comparable to other top systems.

Table 3
Tulu Results

S.No.	Team Name	Run	mF1	Rank
1	NLPfusion	1	0.820	1
2	MUCS_off	3	0.790	2
3	CoreFour	1	0.780	3
4	Dravidian_decoders	2	0.770	4
5	langTeam	-	0.770	4
6	YenLP_CS	2	0.730	5
7	**IReL@IIT-BHU**	**1**	**0.710**	**6**
8	MUCS_DL	3	0.610	7
9	DravidianDefenders	-	0.550	8
10	Coreminds_offensive	-	0.400	9

Table 4
Kannada Results

S.No.	Team Name	Run	mF1	Rank
1	NLP fusion	run1	0.475	1
2	CoreFour	run3	0.434	2
3	CoreMinds	-	0.433	3
4	**IReL@IIT-BHU**	**run2**	**0.430**	**4**
5	MUCS_Off	run1	0.421	5
6	Dravidian_decoders	run1	0.384	6
7	YenLP_cs	run2	0.375	7
8	MUCS_DL	run1	0.344	8

In the Malayalam task, our system (Run 1) achieved an mF1 score of 0.667, ranking **5th among all participating teams** (Table 5). While the best-performing team achieved 0.778, our result demonstrates that multilingual pre-trained transformers such as XLM-RoBERTa can provide a competitive baseline even for less-resourced Dravidian languages.

For Tamil, our early-stopped model (Run 2) obtained an mF1 score of 0.448, securing the **4th rank** on the leaderboard (Table 6). This result is close to the top-performing systems, where the leading team achieved an mF1 score of 0.465, showing that our method is effective for Tamil code-mixed text as well.

Table 5
Malayalam Results

S.No.	Team Name	Run	mF1	Rank
1	CoreFour	run1	0.778	1
2	NLP fusion	run1	0.774	2
3	YenLP_cs	run3	0.750	3
4	MUCS_offensive	run2	0.712	4
5	**IReL@IIT-BHU**	**run1**	**0.667**	**5**
6	langteam	-	0.511	6
7	Dravidian_defenders	run1	0.430	7
8	MUCS_DL	run3	0.350	8
9	CoreMinds	-	0.233	9
10	Malayalam_lan_tech	-	0.140	10

Table 6
Tamil Results

S.No.	Team Name	Run	mF1	Rank
1	NLP fusion	run2	0.465	1
2	CoreFour	run3	0.461	2
3	MUCS_off	run1	0.452	3
4	**IReL@IIT-BHU**	**run2**	**0.448**	**4**
5	DUCS	run1	0.416	5
6	Dravidian_defenders	-	0.386	6
7	YenLP_cs	run2	0.369	7
8	MUCS_DL	run2	0.350	8
9	CoreMinds	-	0.308	9
10	lang team	-	0.267	10

Across the four languages, our systems consistently ranked within the **top-6**, with the best relative performance achieved for **Kannada (4th rank)** and **Tamil (4th rank)**. The results demonstrate the effectiveness of fine-tuned XLM-RoBERTa for offensive language detection in Dravidian code-mixed settings. Furthermore, we observed that incorporating early stopping generally improved model robustness, although the effect varied across languages. These results suggest that multilingual transformer models can serve as a strong baseline for offensive language identification in under-resourced code-mixed languages.

6. Conclusion

In this study, we tackled the issue of detecting offensive language in code-mixed Dravidian languages, specifically Tulu-English, Tamil-English, Malayalam-English, and Kannada-English. We used fine-tuned multilingual transformers (XLM-RoBERTa) with and without early stopping. Our systems consistently ranked in the top six across all tasks. We achieved 6th place in Tulu (mF1 = 0.710), 5th in Malayalam (mF1 = 0.667), and 4th place in both Kannada (mF1 = 0.430) and Tamil (mF1 = 0.448). These results show that transformer-based models are effective in handling linguistic variety, non-native scripts, and frequent code-switching. They also highlight the potential of multilingual pre-trained models as strong starting points for Dravidian languages. We found that early stopping generally improved training stability and robustness, but the benefits varied by language. Despite these encouraging results, we still face challenges in capturing subtle contextual cues, handling ambiguous or sarcastic expressions, and managing limited annotated data. Future work can examine how to adapt multilingual transformers for specific tasks, use data augmentation techniques to address resource shortages, and implement cross-lingual transfer strategies to take advantage of structural similarities among Dravidian languages. Additionally, incorporating outside knowledge sources and developing explainable methods will be crucial for ensuring transparency and cultural sensitivity in classification. Overall, this project provides valuable resources and benchmarks while creating opportunities for more effective and inclusive offensive language detection in underrepresented multilingual contexts.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] 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.

[2] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on twitter, in: Proceedings of the NAACL Student Research Workshop, Association for Computational Linguistics, San Diego, California, 2016, pp. 88–93. URL: https://aclanthology.org/N16-2013/.

[3] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 11, 2017, pp. 512–515.

[4] G. K. Shahi, T. Mandl, M. M. Shah, Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages, in: Proceedings of the 11th Forum for Information Retrieval Evaluation (FIRE 2019), Kolkata, India, 2019, pp. 1–6.

[5] T. Mandl, S. Modha, G. K. Shahi, A. K. Jaiswal, D. Nandini, D. Patel, P. Majumder, J. Schäfer, Overview of the hasoc track at fire 2020: Hate speech and offensive content identification in indo-european languages, in: CEUR Workshop Proceedings, Forum for Information Retrieval Evaluation, Hyderabad, India, 2020, pp. 87–111. URL: https://ceur-ws.org/Vol-2517/T3-4.pdf.

[6] A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, M. Shrivastava, A dataset of hindi-english code-mixed social media text for hate speech detection, in: Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media, Association for Computational Linguistics, 2018, pp. 36–41. URL: https://aclanthology.org/W18-5118/.

[7] P. Mathur, R. Sawhney, M. Ayyar, R. Shah, Did you offend me? classification of offensive tweets in hinglish language, in: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), Association for Computational Linguistics, 2018, pp. 138–148. URL: https://aclanthology.org/W18-5118/.

[8] T. Ghosal, A. Das, R. Kumar, K. Jonnalagadda, M. R. Singh, T. Chakraborty, Sentiment analysis of hindi-english code-mixed social media text, in: Proceedings of the 7th Workshop on South Asian Languages and Linguistics (WSAL2019), 2019.

[9] U. Barman, A. Das, J. Silva, S. Bandyopadhyay, Code-mixing: A challenge for language identification in the language of social media, in: Proceedings of The First Workshop on Computational Approaches to Code Switching, Association for Computational Linguistics, 2014, pp. 13–23. URL: https://aclanthology.org/W14-3902/.

[10] B. R. Chakravarthi, T. Mandl, V. Muralidaran, R. Priyadharshini, J. P. McCrae, Overview of the

track on sentiment analysis for dravidian languages at fire 2020, in: Proceedings of the Forum for Information Retrieval Evaluation, 2020, pp. 21–27.

[11] B. R. Chakravarthi, K. Priyadharshini, V. Muralidaran, J. P. McCrae, M. Arunmozhi, T. Mandl, Findings of the shared task on offensive language identification in dravidian languages at fire 2021, in: Proceedings of the Forum for Information Retrieval Evaluation, 2021, pp. 32–43.

[12] 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.

[13] 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.

[14] 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.

[15] S. Chanda, S. Pal, Irlab@ iitbhu@ dravidian-codemix-fire2020: Sentiment analysis for dravidian languages in code-mixed text, in: Working Notes of FIRE (Forum for Information Retrieval Evaluation), 2020, pp. 535–540.

[16] S. Chanda, S. Ujjwal, S. Das, S. Pal, Fine-tuning pre-trained transformer based model for hate speech and offensive content identification in english indo-aryan and code-mixed (english-hindi) languages, in: Working Notes of FIRE, 2021, pp. 446–458.

[17] S. Chanda, S. Sheth, S. Pal, Coarse and fine-grained conversational hate speech and offensive content identification in code-mixed languages using fine-tuned multilingual embedding, in: Working Notes of FIRE, 2022, pp. 502–512.

[18] S. Chanda, A. Mishra, S. Pal, Sarcasm detection in tamil and malayalam dravidian code-mixed text, in: Working Notes of FIRE, 2023, pp. 336–343.

[19] S. Chanda, S. Pal, The effect of stopword removal on information retrieval for code-mixed data obtained via social media, volume 4, 2023, p. 494.

[20] S. Chanda, K. Tewari, A. Mukherjee, S. Pal, Leveraging chatgpt and xlm-roberta for sarcasm detection in dravidian code-mixed languages, in: Proceedings of FIRE (Working Notes), Forum for Information Retrieval Evaluation, 2024, India, 2024. URL: https://ceur-ws.org/Vol-4054/T4-14.pdf.

[21] S. Chanda, A. Mishra, S. Pal, Advancing language identification in code-mixed tulu texts: Harnessing deep learning techniques, in: Working Notes of FIRE, 2023, pp. 223–230.

[22] S. Chanda, A. Dhaka, S. Pal, Towards safer online spaces: Deep learning for hate speech detection in code-mixed social media conversations, in: Companion Publication of the 16th ACM Web Science Conference, 2024.

[23] S. N, 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.

[24] 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.