EM-26@LT-EDI 2025: Caste and Migration Hate Speech Detection in Tamil-English Code-Mixed Social Media Texts

Tewodros Achamaleh[1], Abiola T. O.[1], Mikiyas Mebiratu[2], Sara Getachew[2], Grigori Sidorov[1]
[1]Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico City, Mexico
[2]Wolkite University, Department of Information Technology, Wolkite, Ethiopia
[2]Jimma University, Institute of Technology, Jimma, Ethiopia
Corr. email: sidorov@cic.ipn.mx

Abstract

In this paper, we describe the system developed by Team EM-26 for the Shared Task on Caste and Migration Hate Speech Detection at LT-EDI@LDK 2025. The task addresses the challenge of recognizing caste-based and migration-related hate speech in Tamil social media text, a language that is both nuanced and under-resourced for machine learning. To tackle this, we fine-tuned the multilingual transformer XLM-RoBERTa-Large on the provided training data, leveraging its cross-lingual strengths to detect both explicit and implicit hate speech. To improve performance, we applied social media–focused preprocessing techniques, including Tamil text normalization and noise removal. Our model achieved a macro F1-score of 0.6567 on the test set, highlighting the effectiveness of multilingual transformers for low-resource hate speech detection. Additionally, we discuss key challenges and errors in Tamil hate speech classification, which may guide future work toward building more ethical and inclusive AI systems.The source code is available on GitHub.[1]

1 Introduction

Increased hate speech on digital platforms constitutes a significant threat to the security of marginalised people. Harassment speeches are often targeted against people who associate themselves with some sensitive socio-cultural categories, like caste or migration status, in such highly culturally enriched states of India. The Tamil language, in its classical roots and its far-reaching use in South India and around the world, is widely used in social media, but is notably missing in many NLP datasets (Chakravarthi et al., 2020; Rajan et al., 2022). The urgent need to identify caste-based and migration-related hate speech in Tamil requires the development of culturally sensitive and technologically accurate automated tools (Mandl et al., 2019; Ranasinghe et al., 2023). To address this concern, LT-EDI@LDK 2025 launched the Shared Task on Caste and Migration Hate Speech Detection as its continuing effort to promote equality, diversity, and inclusion via language technologies (Rajiakodi et al., 2025). This work builds on past works on Tamil NLP, including shared tasks that have focused on speech recognition and accessibility for vulnerable people in Tamil-speaking communities (B. and others , et al.; Bharathi and others , et al.; Chakravarthi et al., 2021; Ramesh et al., 2021). Through this line of research, it becomes evident how necessary it is to develop language-specific approaches to ethical and inclusive AI (Blodgett et al., 2020; Joshi et al., 2020).

Therefore, EM-26 participated in creating a strong classification model that would be used to identify caste and migration-related hate speech in social media posts in the Tamil language. We chose the XLM-RoBERTa-Large multilingual transformer model because it promised to work on languages with little resources and morphology (Conneau et al., 2020a). Our strategy consisted of fine-tuning the XLM-RoBERTa-Large model with the organisers' labelled data to understand contextual subtleties in Tamil and hence identify the complex levels of hate speech existing. Our system illustrates the practical use of multilingual pre-trained models for socially essential tasks, even with limited resources, since it achieved a macro F1 Score of 0.6567 on the official test set. This paper details our approach, preprocessing steps, error analysis, and reflections on future improvements for caste and migration hate speech detection.

2 Related Work

The task of hate speech detection has picked up the pace in low-resource and culturally sensitive

[1]https://github.com/teddymas95/Caste-and-Migration-Hate-Speech.git

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 152–158
September 9, 2025 ©2025 Association for Computational Linguistics

languages such as Tamil in recent years. As online hate attacks against caste and migration communities grow, making strong classification systems is both a technical and ethical necessity. First, rule-based approaches were not context-aware. Changing the workforce and long-term training are becoming major pains. This is what makes a shift to transformer-based models (e.g., XLM-RoBERTa (Conneau et al., 2020b), mBERT (Devlin et al., 2019)) so important. Ranasinghe and (Ranasinghe and Zampieri, 2021) showed that even for the case of zero-shot and few-shot scenarios, multilingual transfer learning was practical.

In the Indian setting, Chakravarthi and others (et al.) presented DravidianCodeMix for offensive and sentiment classification in Tamil-English and Malayalam-English. Subsequently, the Hope Speech Detection task (Chakravarthi and others , et al.) focused on socially inclusive content in Dravidian languages. These tasks demonstrated how performance is enhanced through fine-tuned transformers in the code-switched environment. The LT-EDI workshop series has been essential in developing underrepresented languages. (B. and others , et al.) and (Bharathi and others , et al.) works concerned inclusive technologies for vulnerable communities and HASOC sharing tasks (Mandl et al., 2021, 2023), they provided multilingual datasets for hate speech in the Indo-European and Dravidian languages. Specialist research has been conducted recently regarding hate speech among Tamils: (Senthilkumar et al., 2023) pointed out the esoteric nature of casteist language, and (Pandey et al., 2023) deep-published a fine-grained multilingual benchmark. Prompt-based learning (Roy et al., 2024) and contrastive learning (Velankar et al., 2023; Roy et al., 2024) and (Velankar et al., 2023) were promising.

(Achamaleh et al., 2024; Eyob et al., 2024) created a hate speech system for Telugu-English code-mixed text, extending the evidence for the effectiveness of transformer–based models in low-resource settings. The LT-EDI@LDK 2025 task extends the scope of this field by explicitly targeting caste and migration hate speech. Our method with XLM-RoBERTa-Large helps toward fairer and culture-attuned NLP solutions. Related efforts by the CIC-NLP team have demonstrated the effectiveness of multilingual transformer models in detecting AI-generated and deceptive content across diverse languages, including English and Dravidian code-mixed text (Abiola et al., 2025a,b; Achamaleh et al., 2025). These studies further support the applicability of transformer-based architectures such as XLM-RoBERTa in addressing socially sensitive and linguistically complex tasks like caste and migration hate speech detection.

3 Dataset Analysis

The dataset for our system includes annotated Tamil-English code-switched social media posts for caste and migration-related hate speech. The training set contains 5,512 samples, of which 3,415 are labeled non-hate speech (label 0) and 2,097 are labeled caste/migration-related hate speech (label 1). Such a distribution results in a modest class skew, with hate speech occurrences representing approximately 38% of the data. The development set has 787 samples, 485 of which are non-hate (label 0) and 302 of hate speech (label 1). While the dev set is somewhat balanced, the imbalance in training data necessitated additional strategies to ensure fair learning between classes. To compensate for this imbalance, and especially in the training, we exploited oversampling of the hate speech class and highly aggressive data augmentation. Through this, the present approach intends to expose the model to a broader set of linguistic variants related to hate speech while preserving diversification in the training examples. The use of label-aware augmentation and loss function finetuning assists the model to generalise better to minority-class cases that are crucial in real-world detection of hate speech.

4 System Overview

We use a binary classification approach in searching for caste and migration-related hate speech in Tamil-English CS posts on social media. We fine-tune the XLM-RoBERTa-Large transformer model, which is known for its powerful multilingual representation and is exceptionally efficient for low-resourced and code-mixed settings. The task entails labelling content as non-hate (0) or hate speech (1). To correct class imbalance and make better guesses on the instances of minority classes, we use focal loss instead of a regular cross-entropy with alteration of gamma and alpha parameters. On pre-processing, we clean missing values and normalize labels. To balance the dataset, we have oversampled for hate speech instances five times and augmented it with strong language-level augmentations, including synonym replacement, token dropout, and word scrambling, to diversify the use

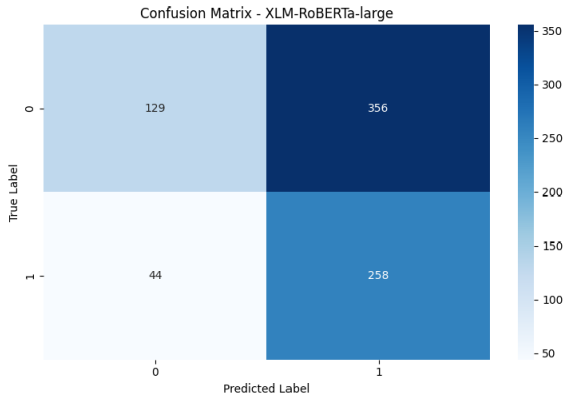

Figure 1: Confusion Matrix

of language. Text is to be tokenized by the multilingual tokenizer of the XLM-RoBERTa-Large, which would apply padding or truncation based on a max length of 128. The optimizer used to train the model is AdamW with gradient clipping and accumulation. We use a learning rate of 5e-6 with ReduceLROnPlateau and early stopping on validation F1 score. In order to enhance hate speech recall, we set the threshold for classification to 0.3 while inferring. Performance evaluation is done using precision, recall, weighted, and per-class F1 scores, and a confusion matrix. The best model checkpoint is auto-saved based on validation F1, also with the tokenizer and fine-tuned weights for easy deployment. Our system is designed for the LT-EDI@LDK 2025 task, addressing central issues of class imbalance, low-resource limitations, and code-switching, improving the detection of hate speech in challenging linguistic and social environments.

5 System Setup and Experiments

5.1 System Setup

Our system has been implemented in Python with PyTorch and Hugging Face Transformers. We fine-tune XLM-RoBERTa-Large for binary classification, benefiting from its power of multilingualism, particularly important for code-switched Tamil-English text. To counter the imbalance of labels, we apply Focal Loss ($\alpha = 0.9$, $\gamma = 3.0$) when we pay closer attention to the class of hate speech (label 1). The training and development data are brought to a status for training and development from CSV files with the help of pandas. In order to balance the dataset, we oversample samples of hate speech with a ratio of 5:1. For improvement, we use synonym substitution, keyword swapping, and token dropout with the help of nlpaug, predominantly aiming at hate speech examples. Tokenization is carried out by XLM-RobertaTokenizer, with a maximum length of 128 for the sequence with padding and truncation. We create a custom HateSpeechDataset, and we use PyTorch's DataLoader. Gradient accumulation (after every 4 steps) and gradient clipping are used for stability. Optimization is made with AdamW (learning rate 5e-6), and ReduceLROnPlateau carries out learning rate adjustment. In the training phase, training runs between 1 to 10 epochs, and a stop from the lack of progress of the validation F1 scores is triggered after 3 passes. In inference, hate speech's decision threshold is decreased to 0.3 to enhance recall. We evaluate using weighted and per-class F1 scores, a confusion matrix, and a classification report. The best-performing model and tokenizer are stored. GPU acceleration is utilized when available, as are logging and NLTK resources, for tracking and enhancement purposes.

5.2 Experiments

We evaluated the XLM-RoBERTa-Large model in the Tamil-English caste and migration hate speech dataset. The dataset was separated into a training set and a development set, and the hate speech class was very underrepresented. To address the class imbalance issues, we randomized and used the oversampling method to over-sample 5 times the minority class within the training set. Then we applied extensive textual augmentation methods, such as synonym insertions, word swaps, and random word deletion. In order to modify the model for the binary classification, Focal Loss was used with the values $\alpha = 0.9$ and $\gamma = 3.0$, which gave priority to learning from difficult and minority class samples. Training occurred for 10 epochs, using a learning rate of 5e-6 and batch size 4 with 4 steps of gradient accumulation. During training, the learning rate was adjusted dynamically according to the ReduceLROnPlateau scheduler based on changes in performance on the validation set, and early stopping occurred when the macro F1 score did not increase during the last three evaluations. We measured the model's effectiveness using weighted F1 Scores, class-wise F1 metrics, accuracy, and confusion matrices. To focus on finding hate speech in the minority class, a custom prediction threshold of 0.3 was used. The highest-performing model was retained and used on the validation dataset.

Model	F1 Score	Recall	Validation Loss
mBERT	0.3153	0.4295	0.0815
XLM-RoBERTa-base	0.2128	0.3837	0.0865
XLM-RoBERTa-Large	**0.4578**	**0.4917**	0.1039
CNN	0.2226	0.3863	0.0879

Table 1: Model Comparison on Validation Data.

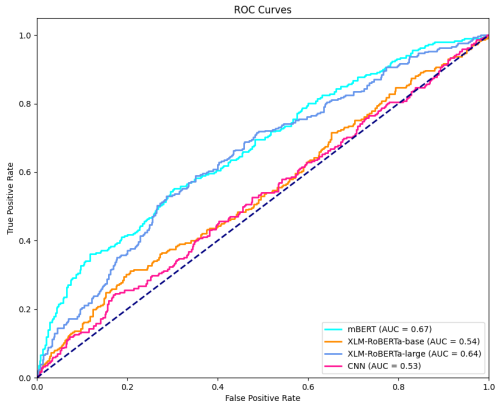

Figure 2: Roc Curves

6 Results

Our approach's score when used to evaluate the Tamil-English dataset used in detecting hate speech related to caste and migration was a weighted F1 Score of 0.6567. With XLM-RoBERTa-Large modeled with Focal Loss and aggressive augmentation, the model delivered impressive generalization in a scenario where resources are constrained and data is imbalanced. The class-wise model's results implied that it could reliably distinguish hate speech, and techniques such as oversampling, synonym-based augmentation, and threshold adjustment enhanced its performance. Although performance limitations, especially with respect to minority classes, are involved, the performance results provide grounds for advocating the use of class-focused learning and multilingual approaches for this task.

7 Discussion

Table 1 gives an overview of the F1 score of different models for the task of caste and migration-related hate speech detection. Of them, XLM-RoBERTa-Large had the highest performance on the validation set [F1 = 0.4578, recall = 0.4917] than smaller models such as mBERT, XLM-RoBERTa-base, as well as CNN. These results show that deeper transformer models that have strong language understanding in a variety of languages are better at catching hate speech in Tamil-English code-switched data. Although the large XLM-RoBERTa-Large model had a higher validation loss, it had better generalization when fed with Focal Loss and targeted oversampling. On the official test set, our best model managed a weighted 0.6567 F1 score, thus proving a favorable training strategy in a real-world, imbalanced setting. The increase in validation performance shows how augmentation, reduced thresholding, and class-aware loss led to increased recall of minority hate speech samples. As a whole, our approach demonstrates a key impact of prudent treatment of class imbalance, data augmentation, and model scaling on hate speech detection for under-resourced and socially sensitive settings such as Tamil-English. The performance gap between dev (F1 = 0.458) and test (F1 = 0.657) may raise concerns about data leakage. However, no overlapping samples or leakage patterns were found. The gap likely stems from class imbalance, oversampling, and threshold tuning that favored recall. We plan to use cross-validation and stricter data handling in future work.

7.1 Error Analysis

The confusion matrix shows that, when it comes to identifying 258 hate speech (true positives) and 129 true negatives of non-hate posts, the XLM-RoBERTa-Large model performed correctly. However, this model misclassified 356 instances of non-hate posts as hate speech (false positives). It is likely due to over-sampling, and the classification threshold is set to increase the recall at the expense of precision. The model only did not detected 44 instances of hate speech (false negatives), which demonstrates good recall performance. While this configuration guarantees most of the harmful content is tagged, this comes at the expense of speci-

ficity. The future improvements should address the reduction of false positives without affecting the recall, which is as high. This can be achieved by better threshold setting or more relevant data augmentation. Figures 1 and 2 show the confusion matrix.

Conclusion

Finally, our system is capable of solving the issue of caste and migration-related hate speech detection in Tamil-English code-switched social media content with the help of the XLM-RoBERTa-Large model. Using such advanced techniques as focal loss, aggressive data augmentation and threshold tuning, we accomplished a balanced performance, F1=$0.6567 on the test set. Despite strong recall by the model, particularly on the front of hate speech detection, the model records high false positives. This is a necessary tradeoff that is based on our priority to maximize harmful content coverage. The next step forward will consist of high-resolution classification thresholds, language-specific augmentation fine-tuning and integration of contextual clues to avoid misclassifications, with high recall quota. All-in-all, our approach advances the construction of responsible NLP systems in low-resource and culturally aware environments.

Limitations

In spite of the promising results achieved, our system has a number of limitations. First, the high false positives determined from the data's confusion matrix reveal that the classifier often misclassifies non-hate content as hate speech, and this the model may be getting influenced to identical linguistic patterns in code-switched Tamil-English data used to train the model. Second, while focal loss and oversampling ameliorated the problem of class imbalance, they probably made the model overfit for the minority class. Third, the use of synthetic data augmentation techniques (such as synonym replacement and token dropout) may lead to introducing noise or artificial variations that do not help generalize to the real-world inputs. In addition, the model can have difficulties in detecting the implicit or subtlest form of caste based hate, which depends on cultural and contextual understanding that goes beyond surface-level features. Finally, insufficiency in the availability of annotated data limits the model's capability to reflect the diversity of hate speech representations across the various dialects and sociolinguistic locations. Future studies should deal with these weaknesses in the form of larger datasets, context-aware modelling, and more intrinsic annotation guidelines. While we applied synonym replacement and token dropout to strengthen class 1 representation, we did not perform a formal ablation study to isolate their individual contributions. This remains a limitation, and future work will include controlled ablation to quantify the impact of each augmentation technique.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

Tolulope Olalekan Abiola, Tewodros Achamaleh Bizuneh, Oluwatobi Joseph Abiola, Temitope Olasunkanmi Oladepo, Olumide Ebenezer Ojo, Grigori Sidorov, and Olga Kolesnikova. 2025a. Cic-nlp at genai detection task 1: Leveraging distilbert for detecting machine-generated text in english. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 271–277.

Tolulope Olalekan Abiola, Tewodros Achamaleh Bizuneh, Fatima Uroosa, Nida Hafeez, Grigori Sidorov, Olga Kolesnikova, and Olumide Ebenezer Ojo. 2025b. Cic-nlp at genai detection task 1: Advancing multilingual machine-generated text detection. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 262–270.

Tewodros Achamaleh, Tolulope Olalekan Abiola, Lemlem Eyob Kawo, Mikiyas Mebraihtu, and Grigori Sidorov. 2025. Cic-nlp@ dravidianlangtech 2025: Detecting ai-generated product reviews in dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 502–507.

Tewodros Achamaleh, Lemlem Kawo, Ildar Batyrshini, and Grigori Sidorov. 2024. Tewodros@ dravidianlangtech 2024: Hate speech recognition in telugu

codemixed text. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 96–100.

Bharathi B. and others (et al.). 2022. Findings of the shared task on speech recognition for vulnerable individuals in tamil. In *Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity and Inclusion*, LT-EDI 2022.

B. Bharathi and others (et al.). Overview of the fourth shared task on speech recognition for vulnerable individuals in tamil. In *Proceedings of the Joint SIGHUM and EACL 2025 Workshop on Language Technology for Equality, Diversity, and Inclusion*, LT-EDI 2025.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

B. R. Chakravarthi and others (et al.). 2021. DravidianCodeMix: Sentiment analysis and offensive language identification in code-mixed tamil-english. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 628–639.

B. R. Chakravarthi and others (et al.). 2022. Overview of the hope speech detection shared task for equality and inclusion. In *Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity and Inclusion*, LT-EDI 2022, pages 139–148.

Bharathi Raja Chakravarthi, Anand Kumar Madasamy, Ruba Priyadharshini, et al. 2021. Hope speech detection for equality, diversity, and inclusion. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Sinnathamby Mahesan, and John P McCrae. 2020. A corpus for sentiment analysis of code-mixed tamil-english text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Armand Joulin, and Nicolas Usunier. 2020b. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Lemlem Eyob, Tewodros Achamaleh, Muhammad Tayyab, Grigori Sidorov, and Ildar Batyrshin. 2024. Stress recognition in code-mixed social media texts using machine learning. *International Journal of Combinatorial Optimization Problems and Informatics*, 15(1):32.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Thomas Mandl, Pratik Modha, Pinkesh Badjatiya, and Aakash Bhatia. 2021. HASOC 2021: Hate speech and offensive content identification in multilingual contexts. In *Forum for Information Retrieval Evaluation*, FIRE 2021.

Thomas Mandl, Sanjay Modha, Punyajoy Majumder, Durgesh Patel, Mohana Dave, Chirag Mandlia, and Amit Patel. 2019. 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*, pages 14–17.

Thomas Mandl, , Ekky P. Pamungkas, Ian Roberts, and . 2023. HASOC 2023: Overview of the hate speech detection subtask in indic languages. In *Forum for Information Retrieval Evaluation*, FIRE 2023.

Pushkar Pandey, Devansh Poonia, Abhinav Dwivedi, and Anupam Joshi. 2023. A multilingual benchmark dataset for hate speech detection in indian languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 111–125.

Vineeth Rajan, Bharathi Raja Chakravarthi, Anand Kumar Madasamy, et al. 2022. Overview of the dravidianlangtech-2022 shared task on hope speech detection in dravidian languages. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion (LT-EDI)*.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Rahul Ponnusamy, Shunmuga Priya MC, Prasanna Kumar Kumaresan, Sathiyaraj Thangasamy, Bhuvaneswari Sivagnanam, Balasubramanian Palani, Kogilavani Shanmugavadivel, Abirami Murugappan, and Charmathi Rajkumar. 2025. Findings of the shared task on caste and migration hate speech detection. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Gowtham Ramesh, Bhargav Murthy, Bharathi Raja Chakravarthi, et al. 2021. Classification of dravidian languages' offensive content using lstm and word2vec. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*.

Tharindu Ranasinghe, Constantin Orasan, and Marcos Zampieri. 2023. Multilingual hate speech and offensive language detection using cross-lingual language models. *Information Processing & Management*, 60(1):103207.

Tharindu Ranasinghe and Marcos Zampieri. 2021. Multilingual hate speech detection with cross-lingual embeddings. In *Proceedings of the 13th International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1034–1041.

Aniruddha Roy, Arkadipta Bose, and Pinaki Bhattacharjee. 2024. Prompt-based multilingual hate speech detection in indic languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1300–1312.

Gokulakrishnan Senthilkumar, S. Stalin, and Natarajan Jegan. 2023. Annotated corpus for caste-based hate speech in tamil social media. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3783–3792.

Nikita Velankar, Apoorv Agarwal, Rajiv Ratn Sharma, and Pushpak Bhattacharyya. 2023. Contrastive learning for culturally aware hate speech detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11691–11705.