byteSizedLLM@DravidianLangTech 2025: Abusive Tamil and Malayalam Text targeting Women on Social Media Using XLM-RoBERTa and Attention-BiLSTM

Rohith Gowtham Kodali
ASRlytics
Hyderabad, India
rohitkodali@gmail.com

Durga Prasad Manukonda
ASRlytics
Hyderabad, India
mdp0999@gmail.com

Maharajan Pannakkaran
ASRlytics
Hyderabad, India
mahamca.kovai@gmail.com

Abstract

This research investigates abusive comment detection in Tamil and Malayalam, focusing on code-mixed, multilingual social media text. A hybrid Attention BiLSTM-XLM-RoBERTa model was utilized, combining fine-tuned embeddings, sequential dependencies, and attention mechanisms. Despite computational constraints limiting fine-tuning to a subset of the AI4Bharath dataset, the model achieved competitive macro F1-scores, ranking 6th for both Tamil and Malayalam datasets with minor performance differences. The results emphasize the potential of multilingual transformers and the need for further advancements, particularly in addressing linguistic diversity, transliteration complexity, and computational limitations.

1 Introduction

Social media platforms enable communication but are increasingly misused for abuse and harassment. Women often face hateful and threatening comments, reflecting deep-rooted societal biases. This gender-based cyberbullying leads to severe psychological, social, and professional harm. Addressing this issue is vital for creating safer online environments.

The Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media at DravidianLangTech@NAACL 2025[1] seeks to tackle this pressing issue by advancing research on online content moderation. This task focuses on detecting abusive comments targeting women in Tamil and Malayalam, two Dravidian languages predominantly spoken in South India. As low-resource languages in the field of natural language processing (NLP), Tamil and Malayalam present unique challenges for developing robust machine learning models. Furthermore, identifying abusive content in these languages is crucial for empowering marginalized communities and bridging the linguistic gap in content moderation research.

We propose a hybrid model that integrates fine-tuned Tamil and Malayalam XLM-RoBERTa, optimized for transliteration-aware data, with an Attention-BiLSTM classifier. XLM-RoBERTa captures cross-lingual and contextual representations, handling mixed-script inputs, while the Attention-BiLSTM identifies sequential dependencies and highlights key features. This fusion combines transformer-based embeddings and recurrent architectures for effective abuse detection in low-resource languages.

This paper details our methodology, experiments, and results, demonstrating our model's effectiveness in detecting abusive comments. We also discuss challenges encountered and suggest future directions for abusive language detection in low-resource settings.

2 Related Work

Research on detecting Hate, Offensive, and Abusive Speech in CodeMix Dravidian languages like Kannada-English, Malayalam-English, and Tamil-English has grown recently. However, challenges such as linguistic diversity, complex grammar, polysemous words, and limited annotated data persist (Anbukkarasi and Varadhaganapathy, 2023; Chakravarthi et al., 2021c). Shared tasks like DravidianLangTech 2021 and HASOC-DravidianCodeMix (Chakravarthi et al., 2021a,b), alongside annotated datasets (Chakravarthi et al., 2020, 2021b, 2022; Devi, 2021; Jose et al., 2020), have enabled significant advancements in this domain.

Participating teams in these shared tasks have utilized multilingual pre-trained transformers for their contextual understanding and fine-tuning capabilities. For example, Saha et al. (2021) leveraged models like XLM-RoBERTa-large, MuRIL, and In-

[1] https://codalab.lisn.upsaclay.fr/competitions/20701

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 200–205

dicBERT, while Balouchzahi et al. (2021) proposed the COOLI Ensemble model with CountVectors and classifiers such as MLP and XGBoost. Other approaches include handling class imbalance with innovative loss functions (Tula et al., 2021; Vasantharajan and Thayasivam, 2021) and employing strategies like pseudo-labeling, multi-task learning, and selective translation for fine-tuning (Hande et al., 2021a,b; Vasantharajan and Thayasivam, 2021).

Traditional machine learning methods, such as SVMs and Random Forest, have also been explored with feature extraction techniques like TF-IDF (Sivalingam and Thavareesan, 2021). Indic-specific models like IndicBART (Dabre et al., 2022) have shown potential in tasks like translation and summarization. Despite these efforts, there remains no widely recognized pre-trained model for hate speech detection in CodeMix Dravidian languages.

Abusive comment detection in Tamil was a key focus of DravidianLangTech 2022(Priyadharshini et al., 2022), where datasets highlighted challenges in handling code-mixed Tamil-English text. In DravidianLangTech 2023(Bala and Krishnamurthy, 2023), the scope expanded to include both Tamil and Telugu, offering new datasets and benchmarks. These tasks spurred advancements in multilingual transformers, ensemble learning, and strategies for tackling class imbalance and data scarcity, further enriching abusive comment detection research in Dravidian languages.

Despite recent advancements, there remains significant scope to improve existing models and develop new, robust approaches for abusive comment detection in Dravidian languages. Addressing challenges like linguistic diversity, complex code-mixing, and limited annotated data will be critical to advancing this field further

3 Dataset

The goal of this task is to identify whether a given comment contains abusive content or not. The task dataset comprises text in Tamil and Malayalam(Priyadharshini et al., 2022), two low-resource languages spoken in South India. Each comment is annotated with binary labels: **Abusive** and **Non-Abusive**.

3.1 Malayalam Dataset

The Malayalam dataset contains a total of 3562 comments, divided into **2933** training and **629** testing instances. Table 1 provides a detailed breakdown of the dataset statistics for Malayalam.

Label	Train	Test	Total
Abusive	1531	323	1854
Non-Abusive	1402	306	1708
Total	2933	629	3562

Table 1: Dataset statistics for Malayalam, including total counts for each label and split.

3.2 Tamil Dataset

The Tamil dataset contains a total of 3388 comments, with **2790** for training and **598** for testing. Table 2 provides a detailed breakdown of the dataset statistics for Tamil.

Label	Train	Test	Total
Non-Abusive	1424	305	1729
Abusive	1366	293	1659
Total	2790	598	3388

Table 2: Dataset statistics for Tamil, including total counts for each label and split.

4 Methodology

This study employs a hybrid Attention BiLSTM-XLM-RoBERTa model(Liu and Guo, 2019; Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005; Conneau et al., 2019; Manukonda and Kodali, 2025; Kodali et al., 2025; Manukonda and Kodali, 2024a; Kodali and Manukonda, 2024; Manukonda and Kodali, 2024b) to classify abusive and non-abusive comments in Tamil and Malayalam. The architecture, shown in Figure 1, combines fine-tuned XLM-RoBERTa embeddings, a bidirectional LSTM (BiLSTM), and an attention mechanism to effectively extract and process features for binary classification.

4.1 Transliteration aware XLM-RoBERTa Fine-tuning

The TamilXLM-RoBERTa and MalayalamXLM-RoBERTa models were fine-tuned using approximately 300MB of monolingual text from AI4Bharath[2](Kunchukuttan et al., 2020) for each language. To handle the diverse script usage in comments, the **IndicTrans**(Bhat et al., 2015) transliteration tool was employed to create three

[2]https://github.com/AI4Bharat/indicnlp_corpus

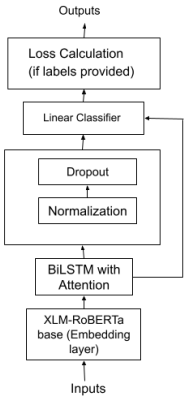

Figure 1: Architecture of the BiLSTM-XLM-RoBERTa Classifier Model.

variations of the dataset: native script text, fully transliterated text in Roman script, and partially transliterated text (20–70% of words transliterated). This approach ensures compatibility with native scripts, Romanized text, and mixed-script text, which are prevalent in social media communication.

4.2 Attention-BiLSTM-XLM-RoBERTa Classifier

The Attention-BiLSTM-XLM-RoBERTa classifier combines three key components: contextual understanding, sequential learning, and attention-based feature selection. XLM-RoBERTa generates contextual embeddings, which are refined by a BiLSTM layer to capture sequential dependencies. An attention mechanism identifies and emphasizes important features, enhancing interpretability. Residual layer normalization and dropout are applied for stability. Finally, a classification layer produces logits, optimized using cross-entropy loss. This hybrid model effectively detects abusive comments in Tamil and Malayalam.

5 Experiment Setup

We fine-tuned TamilXLM-RoBERTa[3] and MalayalamXLM-RoBERTa[4] for multilingual, code-mixed text. Data preprocessing included removing punctuation, HTML tags, and noise, with labels encoded for binary classification. A 90:10 stratified split ensured balanced training and validation sets.

[3] https://huggingface.co/bytesizedllm/TamilXLM_Roberta

[4] https://huggingface.co/bytesizedllm/MalayalamXLM_Roberta

Using **IndicTrans**, three dataset variations were created: native script, fully Romanized, and partially transliterated text, enabling the model to handle mixed-script social media text. Fine-tuned XLM-RoBERTa embeddings were combined with a BiLSTM layer for sequential learning and an attention mechanism for refined feature representation.

The model was trained for 10 epochs with AdamW (2×10^{-5} learning rate, 0.01 weight decay) and a linear scheduler. Gradient clipping (max norm 1.0) stabilized training. Validation after each epoch used accuracy and macro F1-score, with the best model per language selected based on macro F1-score, ensuring robust fine-tuning for detecting abusive comments in Tamil and Malayalam.

Team Name	mF1	Rank
CUET_Agile	0.7883	1
MSM_CUET	0.7873	2
Incepto	0.7864	3
Lowes	0.7824	4
Necto	0.7821	5
byteSizedLLM	**0.7820**	**6**

Table 3: Macro F1 (mF1) scores and ranks of the top 6 performing teams on the Tamil test set.

Team Name	mF1	Rank
Habiba A, G Agila	0.7571	1
CUET_Agile	0.7234	2
CUET_Novice	0.7083	3
Incepto	0.7058	4
Lowes	0.7001	5
byteSizedLLM	**0.6964**	**6**

Table 4: Macro F1 (mF1) scores and ranks of the top 6 performing teams on the Malayalam test set.

6 Results and Discussion

The fine-tuned multilingual model achieved a perplexity of 4.9 for Tamil and 4.1 for Malayalam, demonstrating effective language modeling performance across both languages.

The model performed better on the Tamil dataset, achieving an accuracy of 78% with F1-scores of 0.79 for Abusive and 0.77 for Non-Abusive content. In contrast, the Malayalam dataset had a lower accuracy of 70%, with F1-scores of 0.67 for Abusive and 0.73 for Non-Abusive content. The Tamil dataset exhibited a more balanced precision (0.80)

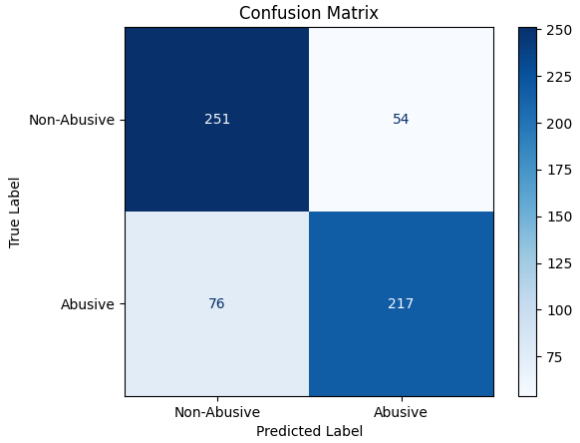

Figure 2: Confusion Matrix for Abusive and Non-Abusive Classification in Tamil

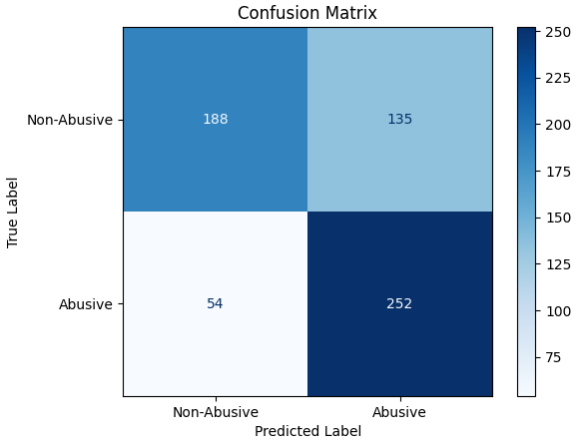

Figure 3: Confusion Matrix for Abusive and Non-Abusive Classification in Malayalam

and recall (0.74) for Non-Abusive content, resulting in more consistent performance.

As shown in Table 3, CUET_Agile achieved the highest mF1 score of 0.7883, followed closely by MSM_CUET (0.7873) and Incepto (0.7864). Our team, **byteSizedLLM**, achieved a Macro F1 score of 0.7820, ranking 6th. This demonstrates the effectiveness of our hybrid Attention BiLSTM-XLM-RoBERTa model in handling Tamil social media data. The close performance among the top teams suggests potential for further improvements through hyperparameter tuning and advanced data augmentation.

For the Malayalam dataset (Table 4), Habiba A, G Agila led with an mF1 score of 0.7571, followed by CUET_Agile (0.7234) and CUET_Novice (0.7083). Our team, **byteSizedLLM**, achieved a Macro F1 score of 0.6964, placing 6th. The lower performance for Malayalam reflects the linguistic diversity and complex code-mixed structures, highlighting the need for better class imbalance handling and transliteration strategies.

The model misclassified 135 abusive instances as non-abusive in Malayalam and 54 in Tamil, indicating better recall for abusive content in Tamil. Malayalam's non-abusive recall was 82%, effectively identifying non-abusive text but missing some abusive cases. Tamil's abusive recall was also 82%, reflecting a bias toward the majority class. Refer to Fig.2 for Tamil and Fig.3 for Malayalam.

The results show the effectiveness of our approach in detecting abusive comments in code-mixed multilingual text. The top team's strong performance highlights the potential of multilingual transformers, though improvements are needed for linguistic nuances and robustness to diverse code-mixed patterns

7 Limitations and Future Work

A key limitation was computational constraints, which restricted fine-tuning TamilXLM-RoBERTa and MalayalamXLM-RoBERTa models to a small subset of the AI4Bharat dataset. This hindered the model's ability to fully utilize the dataset's linguistic diversity and contextual richness.

For Tamil, the narrow performance gap among top teams indicates that transformer-based approaches have matured. However, for Malayalam, challenges like transliteration complexity and limited training data persist. Future work should address computational limits, explore pseudo-labeling and ensemble learning, and integrate external linguistic resources to improve performance.

8 Conclusion

This study demonstrated the potential of hybrid Attention BiLSTM-XLM-RoBERTa models for abusive comment detection in Tamil and Malayalam[5]. Despite computational constraints, our approach achieved competitive results, underscoring the effectiveness of integrating multilingual embeddings with sequential and attention mechanisms.

Future research should further refine these models by leveraging larger datasets, optimizing hyperparameters, and enhancing domain adaptation techniques to improve robustness and generalization.

[5] https://github.com/mdp0999/Abusive-Tamil-and-Malayalam-Text

References

S. Anbukkarasi and S. Varadhaganapathy. 2023. Deep learning-based hate speech detection in code-mixed tamil text. *IETE Journal of Research*, 69(11):7893–7898.

Abhinaba Bala and Parameswari Krishnamurthy. 2023. AbhiPaw@ DravidianLangTech: Abusive comment detection in Tamil and Telugu using logistic regression. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 231–234, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Fazlourrahman Balouchzahi, Aparna B K, and H L Shashirekha. 2021. MUCS@DravidianLangTech-EACL2021:COOLI-code-mixing offensive language identification. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 323–329, Kyiv. Association for Computational Linguistics.

Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2015. Iiit-h system submission for fire2014 shared task on transliterated search. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '14, pages 48–53, New York, NY, USA. ACM.

Bharathi Raja Chakravarthi, Mariappan Anandkumar, John P. McCrae, Bhavukam Premjith, K. P. Soman, and Thomas Mandl. 2020. Overview of the track on hasoc-offensive language identification-dravidiancodemix. In *Fire*.

Bharathi Raja Chakravarthi, Dhivya Chinnappa, Ruba Priyadharshini, Anand Kumar Madasamy, Sangeetha Sivanesan, Subalalitha Chinnaudayar Navaneethakrishnan, Sajeetha Thavareesan, Dhanalakshmi Vadivel, Rahul Ponnusamy, and Prasanna Kumar Kumaresan. 2021a. Developing successful shared tasks on offensive language identification for dravidian languages. *Preprint*, arXiv:2111.03375.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John P. McCrae, and Elizabeth Sherly. 2021b. 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*, pages 133–145, Kyiv. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2022. Dravidiancodemix: sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. *Language Resources and Evaluation*, 56(3):765–806.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Sajeetha Thavareesan, Dhivya Chinnappa, Durairaj Thenmozhi, Elizabeth Sherly, John P. McCrae, Adeep Hande, Rahul Ponnusamy, Shubhanker Banerjee, and Charangan Vasantharajan. 2021c. Findings of the sentiment analysis of dravidian languages in code-mixed text. *Preprint*, arXiv:2111.09811.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

Sobha Lalitha Devi. 2021. Anaphora resolution from social media text in indian languages (socanares-il)-overview. In *Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '20, page 9–13, New York, NY, USA. Association for Computing Machinery.

A. Graves and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052 vol. 4.

Adeep Hande, Siddhanth U Hegde, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021a. Benchmarking multi-task learning for sentiment analysis and offensive language identification in under-resourced dravidian languages. *Preprint*, arXiv:2108.03867.

Adeep Hande, Karthik Puranik, Konthala Yasaswini, Ruba Priyadharshini, Sajeetha Thavareesan, Anbukkarasi Sampath, Kogilavani Shanmugavadivel, Durairaj Thenmozhi, and Bharathi Raja Chakravarthi. 2021b. Offensive language identification in low-resourced code-mixed dravidian languages using pseudo-labeling. *Preprint*, arXiv:2108.12177.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Navya Jose, Bharathi Raja Chakravarthi, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2020. A survey of current datasets for code-switching research. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 136–141.

Rohith Kodali and Durga Manukonda. 2024. byteSizedLLM@DravidianLangTech 2024: Fake news detection in Dravidian languages - unleashing the power of custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the*

Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 79–84, St. Julian's, Malta. Association for Computational Linguistics.

Rohith Gowtham Kodali, Durga Prasad Manukonda, and Daniel Iglesias. 2025. byteSizedLLM@NLU of Devanagari script languages 2025: Hate speech detection and target identification using customized attention BiLSTM and XLM-RoBERTa base embeddings. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 242–247, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

Gang Liu and Jiabao Guo. 2019. Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337:325–338.

Durga Manukonda and Rohith Kodali. 2024a. byteLLM@LT-EDI-2024: Homophobia/transphobia detection in social media comments - custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 157–163, St. Julian's, Malta. Association for Computational Linguistics.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2024b. Enhancing multilingual natural language processing with custom subword tokenization: Subword2vec and bilstm integration for lightweight and streamlined approaches. In *2024 6th International Conference on Natural Language Processing (ICNLP)*, pages 366–371.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2025. byteSizedLLM@NLU of Devanagari script languages 2025: Language identification using customized attention BiLSTM and XLM-RoBERTa base embeddings. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 248–252, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and booktitle = Kumaresan, Prasanna Kumar". Findings of the shared task on Abusive Comment Detection in Tamil and Telugu.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Debjoy Saha, Naman Paharia, Debajit Chakraborty, Punyajoy Saha, and Animesh Mukherjee. 2021. Hatealert@DravidianLangTech-EACL2021: Ensembling strategies for transformer-based offensive language detection. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 270–276, Kyiv. Association for Computational Linguistics.

Disne Sivalingam and Sajeetha Thavareesan. 2021. OffTamil@DravideanLangTech-EASL2021: Offensive language identification in Tamil text. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 346–351, Kyiv. Association for Computational Linguistics.

Debapriya Tula, Prathyush Potluri, Shreyas Ms, Sumanth Doddapaneni, Pranjal Sahu, Rohan Sukumaran, and Parth Patwa. 2021. Bitions@DravidianLangTech-EACL2021: Ensemble of multilingual language models with pseudo labeling for offence detection in Dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 291–299, Kyiv. Association for Computational Linguistics.

Charangan Vasantharajan and Uthayasanker Thayasivam. 2021. Towards offensive language identification for tamil code-mixed youtube comments and posts. *SN Computer Science*, 3(1).