byteSizedLLM@DravidianLangTech 2025: Detecting AI-Generated Product Reviews in Dravidian Languages 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 study presents a hybrid model integrating TamilXLM-RoBERTa and MalayalamXLM-RoBERTa with BiLSTM and attention mechanisms to classify AI-generated and human-written product reviews in Tamil and Malayalam. The model employs a transliteration-based fine-tuning strategy, effectively handling native, Romanized, and mixed-script text. Despite being trained on a relatively small portion of data, our approach demonstrates strong performance in distinguishing AI-generated content, achieving competitive macro F1 scores in the DravidianLangTech 2025 shared task. The proposed method showcases the effectiveness of multilingual transformers and hybrid architectures in tackling low-resource language challenges.

1 Introduction

The rapid advancement of artificial intelligence (AI) has significantly transformed natural language processing (NLP) and content generation. While these developments enhance text-based applications, they also facilitate the proliferation of AI-generated content, posing challenges to domains that rely on textual authenticity, such as online product reviews. The increasing sophistication of synthetic text generation necessitates effective detection mechanisms to preserve content credibility (Ben Jabeur et al., 2023).

To address this issue, the Shared Task on Detecting AI-Generated Product Reviews in Dravidian Languages, organized as part of DravidianLangTech 2025[1], focuses on detecting synthetic content in Malayalam and Tamil (Premjith et al., 2025). While extensive research exists for high-resource languages like English, AI-generated text detection in Dravidian languages remains underexplored. The complex morphological structures, agglutinative nature, and unique syntactic properties of these languages present additional challenges.

We propose a hybrid model combining fine-tuned, transliteration-aware XLM-RoBERTa (Conneau et al., 2019) with an Attention-BiLSTM (Liu and Guo, 2019) classifier. XLM-RoBERTa captures linguistic nuances through robust cross-lingual representation learning, while the BiLSTM layer, enhanced with attention mechanisms, improves sequential dependency learning and feature prioritization. This integration of transformer-based architectures with recurrent neural networks enhances the detection of AI-generated content.

This paper details our methodology, experimental setup, and results, demonstrating the effectiveness of our approach. We also discuss the challenges of detecting AI-generated text in Dravidian languages and explore future directions for improving content authenticity verification in low-resource linguistic settings.

2 Related Work

The rise of generative AI has raised concerns about its misuse in creating deceptive content like fake product reviews. Luo et al. (2023); Ben Jabeur et al. (2023) proposed a supervised learning framework using statistical theories to detect AI-generated reviews by identifying outliers in feature distributions. Similarly, Gupta et al. (2024) reviewed advancements in fake review detection, emphasizing hybrid frameworks and challenges in detecting AI-generated content.

AI-generated reviews typically feature two categories: novel features from large language models (LLMs) and traditional linguistic features. LLM-generated text tends to be more readable but templated due to predictive word selection, while human-authored text shows more unpredictability and lexical diversity (Guo et al., 2023). Detection metrics like perplexity and burstiness, used in tools

[1] https://codalab.lisn.upsaclay.fr/competitions/20700

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 194–199
May 3, 2025

like GPTZero (Tian and Cui, 2023), measure text randomness and aid in identifying AI-generated content (Cai and Cui, 2023; Liang et al., 2023).

Traditional linguistic features, including sentiment polarity, adjective ratios, and reviewer behavior, have been effective in detecting fake reviews (Yin et al., 2021; Kumar et al., 2022). However, integrating LLM-based and traditional features remains underexplored.

Detecting AI-generated reviews in Malayalam and Tamil is challenging due to their complex morphology and syntax. This work addresses the gap by integrating LLM-based and linguistic features for better detection in low-resource languages.

3 Dataset

This study employs a dataset for detecting AI-generated product reviews in Tamil and Malayalam (Premjith et al., 2025). The task dataset is labeled into two categories: **AI-generated** and **HUMAN-written** reviews. The statistics for both languages are presented in Tables 1 and 2.

Label	Train	Test	Total
AI	405	48	453
HUMAN	403	52	455
Total	808	100	908

Table 1: Tamil dataset distribution across training and test splits.

Label	Train	Test	Total
HUMAN	400	100	500
AI	400	100	500
Total	800	200	1000

Table 2: Malayalam dataset distribution across training and test splits.

The Tamil dataset consists of 908 reviews, with 808 for training and 100 for testing, maintaining a balanced distribution between AI-generated and HUMAN-written reviews. Similarly, the Malayalam dataset comprises 1,000 reviews, with 800 for training and 200 for testing, equally split across both categories. Both datasets follow a **90:10 ratio** for training and development, ensuring stratified splits for robust evaluation.

4 Methodology

This study employs a hybrid Attention BiLSTM-XLM-RoBERTa model (Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005; Kodali et al., 2025; Manukonda and Kodali, 2025, 2024a; Kodali and Manukonda, 2024; Manukonda and Kodali, 2024b) to classify AI-generated and HUMAN-written product reviews in Tamil and Malayalam. The architecture, shown in Figure 1, combines the strengths of fine-tuned XLM-RoBERTa embeddings, a bidirectional LSTM (BiLSTM), and an attention mechanism to effectively extract and process features for classification.

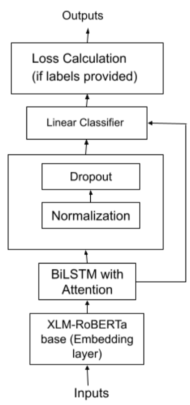

Figure 1: Architecture of the BiLSTM-XLM-RoBERTa Classifier Model.

4.1 Transliteration aware XLM-RoBERTa Fine-tuning

The TamilXLM-RoBERTa[2] and MalayalamXLM-RoBERTa[3] models were fine-tuned using a transliteration strategy with the **IndicTrans** tool (Bhat et al., 2015), leveraging approximately 300MB of monolingual text from AI4Bharath (Kunchukuttan et al., 2020) for each language. The dataset included three variations: native script text, fully transliterated text in Roman script, and partially transliterated text where 20–70% of words were transliterated. This approach enables the model to handle native scripts, Romanized text, and mixed-script text, which are common in social media communication.

4.2 Attention-BiLSTM-XLM-RoBERTa Classifier

The Attention-BiLSTM-XLM-RoBERTa classifier integrates contextual embeddings, sequential modeling, and attention-based feature selection. The

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

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

input sequence is processed by a fine-tuned XLM-RoBERTa model to generate contextual embeddings:

$$\mathbf{X} = \textbf{XLMRoBERTa}(input\_ids, att\_mask) \tag{1}$$

These embeddings are passed through a BiLSTM layer, capturing sequential dependencies by concatenating forward and backward hidden states:

$$\mathbf{H}_t = [\mathbf{H}_{fwd,t}; \mathbf{H}_{bwd,t}] \tag{2}$$

An attention mechanism assigns importance weights to hidden states:

$$\mathbf{a}_t = \tanh(\mathbf{W}_{att} \cdot \mathbf{H}_t), \quad \alpha_t = \frac{\exp(\mathbf{a}_t)}{\sum_{t=1}^{T} \exp(\mathbf{a}_t)} \tag{3}$$

The weighted sum of hidden states forms the attended representation:

$$\mathbf{H}_{attended} = \sum_{t=1}^{T} \alpha_t \cdot \mathbf{H}_t \tag{4}$$

Layer normalization and dropout stabilize training:

$$\mathbf{H}_{dropout} = Dropout(LayerNorm(\mathbf{H}_{attended})) \tag{5}$$

Finally, a classification layer produces logits:

$$\textbf{logits} = \mathbf{W}_{cls} \cdot \mathbf{H}_{dropout} + \mathbf{b}_{cls} \tag{6}$$

Training is optimized using the cross-entropy loss function:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{7}$$

This architecture effectively combines XLM-RoBERTa embeddings, BiLSTM for sequential learning, and attention for key feature selection, enhancing multi-label classification performance.

5 Experiment Setup

We fine-tuned Tamil XLM-RoBERTa and Malayalam XLM-RoBERTa for monolingual and multilingual text classification. The datasets were processed using a data preprocessing pipeline, and labels were encoded as integers for multi-class classification. The data was split into 90% training and 10% validation using a stratified approach.

The fine-tuned XLM-RoBERTa embeddings were integrated with a BiLSTM layer with a hidden size of 512, 3 LSTM layers, and a dropout probability of 0.3. An attention mechanism was added to refine the feature representation. The model was trained for 10 epochs using the AdamW optimizer with a learning rate of 2.5×10^{-5}, weight decay of 0.01, and a linear learning rate scheduler. A batch size of 16 was used, and gradient clipping with a maximum norm of 1.0 was applied for stability.

Validation used accuracy and macro F1-score per epoch, saving the best model for each language to ensure effective fine-tuning for detecting AI-generated reviews in Tamil and Malayalam.

6 Results and Discussion

Team Name	mF1	Rank
KaamKro	0.9199	1
Nitiz - StarAtNyte	0.915	2
Three_Musketeers	0.915	2
SSNTrio	0.9147	3
byteSizedLLM	**0.9**	**4**
Lowes	0.9	4

Table 3: Macro F1 (mF1) scores and ranks of the top 4 performing teams on the Malayalam test set.

Team Name	mF1	Rank
KEC_AI_NLP	0.97	1
CUET_NLP_FiniteInfinity	0.97	1
CIC-NLP	0.96	2
KaamKro	0.95	3
KEC-Elite-Analysts	0.9499	4
byteSizedLLM	**0.94**	**5**

Table 4: Macro F1 (mF1) scores and ranks of the top 5 performing teams on the Tamil test set.

Our experiments demonstrate the effectiveness of the fine-tuned TamilXLM-RoBERTa and MalayalamXLM-RoBERTa models in classifying AI-generated and HUMAN-written product reviews[4]. The perplexity scores achieved by the models underline their capability to adapt to the linguistic nuances of the respective languages, with the Malayalam model achieving a perplexity of 4.1 and the Tamil model achieving a perplexity of 4.9.

Table 3 highlights the performance of the top-performing teams on the Malayalam test set. Our

[4] https://github.com/mdp0999/Detecting-AI-generated-product-reviews

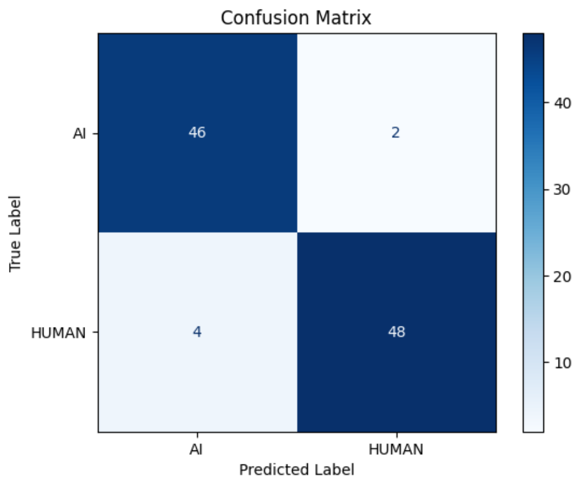

Figure 2: Confusion Matrix for Tamil AI-Generated vs. Human-Written Review Classification

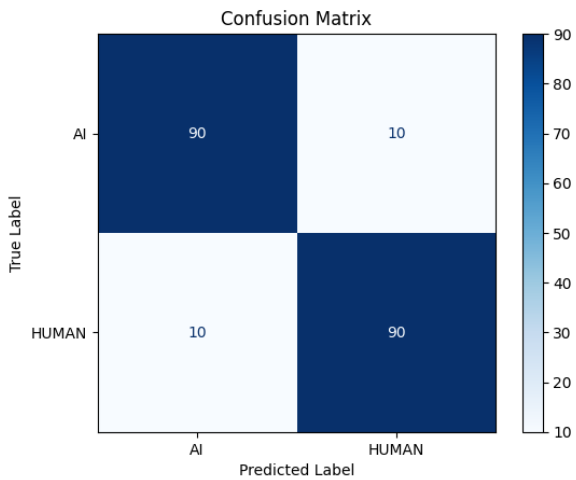

Figure 3: Confusion Matrix for Malayalam AI-Generated vs. Human-Written Review Classification

team, **byteSizedLLM**, secured a shared **4th place** with a Macro F1 (mF1) score of **0.9**. This outcome reflects the strength of our hybrid architecture, which integrates fine-tuned XLM-RoBERTa embeddings with BiLSTM layers and attention mechanisms to address the complexities of Malayalam text effectively.

For the Tamil test set, as summarized in Table 4, our team achieved an mF1 score of **0.94**, placing **5th** among the top teams. The slightly higher perplexity for Tamil indicates challenges in modeling the language, potentially due to its linguistic structure or the dataset's characteristics. Nonetheless, the results validate the robustness of our transliteration-based fine-tuning strategy in managing native, Romanized, and mixed-script text.

The confusion matrices reveal that the model achieves balanced performance across both classes, with very few false positives and false negatives. However, the slightly lower recall for the HUMAN class in Tamil suggests that the model may occasionally misclassify human-written reviews as AI-generated, warranting further optimization. For better understanding, please refer to Fig.2 for Tamil and Fig.3 for Malayalam.

6.1 Limitations and Future Work

Our models were fine-tuned on a limited portion of the available datasets (approximately 300MB per language), constrained by computational resources. This limited dataset size may have restricted the models' ability to fully exploit the linguistic diversity of Tamil and Malayalam. Despite these constraints, the models demonstrated strong performance, but further improvements could be achieved with larger datasets and enhanced computational capabilities.

Future work will focus on scaling the fine-tuning process to utilize more extensive datasets, enabling deeper language modeling. Additionally, adopting advanced strategies such as dynamic data augmentation, multi-task learning, and incorporating more sophisticated preprocessing techniques could further refine model performance. These enhancements aim to reduce perplexity and boost classification accuracy for AI-generated product reviews across multilingual contexts.

7 Conclusion

This study successfully fine-tuned TamilXLM-RoBERTa and MalayalamXLM-RoBERTa models to classify AI-generated and HUMAN-written product reviews. Despite computational constraints limiting the dataset size, the models delivered strong performance, achieving Macro F1 scores of **0.94** for Tamil and **0.9** for Malayalam, ranking among the top teams in their respective tasks. The transliteration-based fine-tuning strategy, combined with a robust hybrid architecture, proved effective in processing diverse scripts, including native, Romanized, and mixed-script text. Remarkably, although the training data was monolingual, the approach demonstrated an ability to generalize to multilingual and mixed-script scenarios, making it highly adaptable for real-world multilingual text classification challenges.

References

Sami Ben Jabeur, Hossein Ballouk, Wissal Ben Arfi, and Jean-Michel Sahut. 2023. Artificial intelligence applications in fake review detection: Bibliometric analysis and future avenues for research. *Journal of Business Research*, 158:113631.

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.

Shuyang Cai and Wanyun Cui. 2023. Evade chatgpt detectors via a single space. *Preprint*, arXiv:2307.02599.

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.

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.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *Preprint*, arXiv:2301.07597.

Richa Gupta, Vinita Jindal, and Indu Kashyap. 2024. Recent state-of-the-art of fake review detection: a comprehensive review. *The Knowledge Engineering Review*, 39:e8.

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

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.

Ajay Kumar, Ram Gopal, Ravi Shankar, and Kim Tan. 2022. Fraudulent review detection model focusing on emotional expressions and explicit aspects: investigating the potential of feature engineering. *Decision Support Systems*, 155:113728.

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*.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. Gpt detectors are biased against non-native english writers. *Preprint*, arXiv:2304.02819.

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

Jiwei Luo, Jian Luo, Guofang Nan, and Dahui Li. 2023. Fake review detection system for online e-commerce platforms: A supervised general mixed probability approach. *Decision Support Systems*, 175:114045.

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.

B Premjith, Nandhini K, Bharathi Raja Chakravarthi, Thenmozhi Durairaj, Balasubramanian Palani, and Kumaresan Prasanna Kumar Thavareesan, Sajeetha. 2025. Overview of the Shared Task on Detecting AI Generated Product Reviews in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Edward Tian and Alexander Cui. 2023. Gptzero: Towards detection of ai-generated text using zero-shot and supervised methods".

Chunyong Yin, Haoqi Cuan, Yuhang Zhu, and Zhichao Yin. 2021. Improved fake reviews detection model based on vertical ensemble tri-training and active learning. *ACM Trans. Intell. Syst. Technol.*, 12:33:1–33:19.