NLP_goats@DravidianLangTech 2025: Detecting AI-Written Reviews for Consumer Trust

Srihari V K
Sri Sivasubramaniya Nadar College of Engineering
srihari2210434@ssn.edu.in

Vijay Karthick Vaidyanathan
Sri Sivasubramaniya Nadar College of Engineering
vijaykarthick2210930@ssn.edu.in

Mugilkrishna D U
Sri Sivasubramaniya Nadar College of Engineering
mugilkrishna2210314@@ssn.edu.in

Durairaj Thenmozhi
Sri Sivasubramaniya Nadar College of Engineering
theni_d@ssn.edu.in

Abstract

The rise of AI-generated content has introduced challenges in distinguishing machine-generated text from human-written text, particularly in low-resource languages. Identifying artificial intelligence (AI)-based reviews is important to preserve trust and authenticity on online platforms. The Shared Task on Detecting AI-Generated Product Reviews in Dravidian languages deals with detecting AI-generated and human-written reviews in Tamil and Malayalam. To solve this problem, we specifically fine-tuned mBERT for binary classification. Our system achieved 10th place in Tamil with a macro F1-score of 0.90 and 28th place in Malayalam with a macro F1-score of 0.68, as the NAACL 2025 organizers reported. The findings demonstrate the complexity of separating AI-derived text from human-authored writing, with a call for continued advances in detection methods. The fine-tuned mBERT model achieved high performance for Tamil, macro F1-score of 0.90 and a score of 0.68 Malayalam. This highlights that some inherent challenges still persist in processing low-resource languages and further language-specific enhancements are needed.

1 Introduction

E-commerce has transformed consumer behaviour, enabling them to share product experiences through reviews on platforms like Amazon and Flipkart. However, the increasing use of AI-generated reviews raises concerns about authenticity, trust, and misinformation in digital markets (Li et al., 2022). AI-powered reviews can manipulate ratings, deceive consumers, and undermine trust, making distinguishing between human- and machine-generated content difficult. Advances in AI text generation further exacerbate this issue, necessitating effective detection mechanisms (Zellers et al., 2019).

Detecting AI-generated reviews is challenging across languages due to the sophistication of generative models and the scarcity of high-quality labelled data. The problem is even more severe in underrepresented Dravidian languages like Tamil, Malayalam, and Telugu, which lack sufficient computational resources and annotated datasets. Their complex linguistic structures, deep morphology, and code-mixing further complicate detection. Reliable AI detection strategies are crucial to maintaining trust in online marketplaces. This study also opens the door for future work, which might explore alternative architectures and larger datasets to overcome current limitations.

This shared task addresses these challenges with two subtasks: Task 1 differentiates human-written reviews from AI-generated reviews in a given dataset, while Task 2 identifies AI-written reviews particularly in Tamil and Malayalam.

This paper is structured as follows: Section 2 reviews prior work on AI-based text detection in low-resource languages, Section 3 details the task descriptions, and Section 4 outlines the methodology, including data preprocessing and model selection and additional model implementations. Section 5 presents experimental results, followed by error analysis in Section 6. Finally, Section 7 concludes with key findings and contributions.

Our study focuses on improving AI-generated review detection in low-resource languages. We aim to develop reliable methods for identifying fake reviews using models like M-BERT, XLM-R and classifiers like Naïve Bayes, contributing to advancements in NLP and AI-generated content detection.

For implementation, please refer to this GitHub repository (srihari2704).

2 Related Work

Detecting AI-generated text is still an open problem because models such as GPT-4 and ChatGPT

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 444–448
May 3, 2025 ©2025 Association for Computational Linguistics

generate increasingly human-like text (Brown et al., 2020). Though detection based on linguistic heuristics and statistical approaches once dominated, nowadays, deep learning and transformers are preferred. It is incredibly challenging in product review contexts, where AI-based contents replicate human styles, requiring further effort for detection (Ippolito et al., 2020). With the increasing popularity of AI-aided review generation, efficient detection methods are essential to guarantee the genuineness of online platforms (Zhang et al., 2020).

In (Fagni et al., 2021), the author investigated AI-produced fabricated content in online reviews, presenting a dataset TweepFake, which includes human-natively and AI-infused product-based and social media reviews. They compared sentiment, coherence, and repetition between AI and human written reviews. Their experiments demonstrated that optimized transformer-based models (e.g., BERT, Roberta) helped traditional model classifiers by using attention-based mechanisms to identify inconsistencies in syntactic patterns between AI-generated reviews. The study concluded that detection accuracy improves significantly when classifiers are trained on domain-specific AI-generated review data rather than general-purpose datasets. Work in artificial product review generation has also yielded clues to enhancing detection.

(Li et al., 2020) studied methods for generating deceptive reviews using AI models and analyzed their effectiveness in fooling human evaluators. Their findings demonstrated that state-of-the-art generative models could produce realistic yet generic-sounding reviews, often lacking the nuanced storytelling in human-authored content. These findings indicate that detecting AI-generated reviews should target linguistic patterns, i.e., illogical coherence, redundant sentences, and high sentiment repetition.

The paper (Wu et al., 2021) also investigated the creation and recognition of AI-generated reviews and their implications for e-commerce synthetic content. Their study examined how AI-generated reviews impact consumer trust and purchasing decisions, underscoring the need for detection frameworks that incorporate both linguistic and behavioural features. Their study suggested hybrid models that integrate BERT-based classification with user behaviour analysis and found that allowing the use of metadata–that is, metadata about review times and user actions–to influence the detection system could significantly improve

ID	DATA	LABEL
TAM_HUAI_TR_001	இந்த சோப்பின் ம	AI
TAM_HUAI_TR_002	தோலை நன்கு சுத்	AI
TAM_HUAI_TR_003	இதைப் பயன்படுத்	AI
TAM_HUAI_TR_004	இந்த சோப்பில் இ	AI
TAM_HUAI_TR_005	சிறிது சோப்பு போ	AI

Figure 1: Dataset for Tamil

MAL_HUAI_TR_398	പേരാമംഗലംരാമ	HUMAN
MAL_HUAI_TR_399	കപ്പയും , മീൻ ക	HUMAN
MAL_HUAI_TR_400	നന്നായിട്ടുണ്ട്. ചോ	HUMAN
MAL_HUAI_TR_401	ഞാൻ ഈ ഫോസ്	AI

Figure 2: Dataset for Malayalam

detection performance.

Previous works have demonstrated the effectiveness of transformer models such as mBERT in cross-lingual tasks. However, alternative models like XLM-R have great potential, given low-resource settings, acting as an important direction for future comparisons.

In addition to transformer-based approaches, models such as Logistic Regression have also been utilized. Although these studies have contributed immensely to the detection of AI-written product reviews, nothing is available in Dravidian languages such as Tamil and Malayalam. These languages have high morphological complexity, which makes adapting to new detection models somewhat difficult. The present study seeks to close this gap by training tailor-made transformer-based models for AI-driven review detection against Dravidian languages to provide more stable e-commerce and digital platforms.

3 Task Description

The task aims to detect AI-generated product reviews in Dravidian languages like Malayalam and Tamil, ensuring authenticity for consumer trust. Participants develop models to distinguish human reviews from AI-generated reviews, using data sets from previous studies (Premjith et al., 2025). Figures 1 and 2 show the Tamil and Malayalam datasets.

4 Methodology

Classifying AI-generated and human-written reviews in Tamil and Malayalam is challenging due to their complex linguistic structures. The model

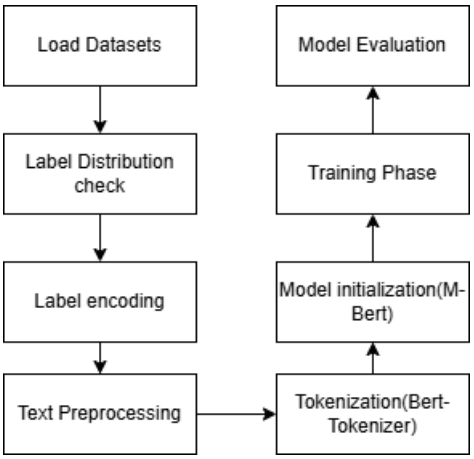

Figure 3: Flowchart representing the process of detecting AI-generated vs human-written reviews

must capture linguistic variation, context, and style for accurate binary classification. The goal is to ensure content authenticity and enhance the reliability of online reviews in Dravidian languages.

Figure 3 represents the overall process for classifying the AI-generated product review.

4.1 Data Preprocessing

Effective preprocessing is essential for improving model performance and distinguishing AI-generated from human-written product reviews in Tamil and Malayalam. The raw dataset undergoes several preprocessing steps to clean and standardize the text.

First, we address missing and inconsistent data by replacing missing text entries with an empty string and mapping non-standardized labels to "AI" for machine-generated reviews and "HUMAN" for human-written reviews.

Next, we clean the text by removing special characters, punctuation, unrelated symbols, non-Tamil/Malayalam symbols, and numerical digits, ensuring linguistic consistency. The text is then converted to lowercase, and redundant spaces are normalized.

After cleaning, the text is tokenized using a BERT-based multilingual tokenizer, resulting in sequences of subword tokens. Shorter sequences are padded to maintain a fixed input length of 256 tokens, while longer sequences are truncated.

Label encoding converts the categorical labels into numerical values, assigning 0 to "AI" and 1 to "HUMAN" for effective processing in a supervised learning context.

The dataset is further stratified into training (80) and testing (20) subsets. This way, AI-generated and human-written reviews are proportionally represented in each subgroup, avoiding class imbalance problems. The label distribution for both AI-generated and human-written reviews in the Tamil and Malayalam datasets is presented in Table 2 and Table 1, respectively. These preprocessing techniques help optimize the dataset to train a robust classification model that can distinguish between AI-generated and human-authored product reviews in Dravidian languages.

Label	Count
AI	405
HUMAN	403

Table 1: Label Distribution in the Malayalam Dataset

Label	Count
AI	410
HUMAN	398

Table 2: Label Distribution in the Tamil Dataset

4.2 Model Evaluation

Recent advancements in Natural Language Processing (NLP) have demonstrated the remarkable capabilities of transformer-based models, especially in tasks involving cross-linguistic text classification. Among these models, mBERT, and XLM-R have shown significant promise in capturing complex contextual information across languages, making them highly effective for text classification tasks such as detecting AI-generated product reviews.

The mBERT model has been pre-trained on a diverse, multilingual corpus, including Dravidian languages like Tamil and Malayalam, which are often considered low-resource languages in the context of NLP (Pires et al., 2019). This extensive pre-training enables mBERT to handle many linguistic features and language structures, which results in a strong performance on tasks with limited annotated data.

XLM-R (Cross-lingual Language Model - RoBERTa) is a strong multilingual model from the RoBERTa architecture, specially designed to manage multiple languages with cross-lingual pre-training (Conneau et al., 2020). It is highly suited for tasks involving generalization over various language structures, such as abusive comment classification.

Also, a Logistic Regression classifier was utilized with TF-IDF features from the preprocessed

text. This method exploits statistical patterns within the text in which word frequency and significance are employed to predict the reviews as being either human or AI-written. Although being straightforward relative to more complex models such as M-BERT and XLM-R, Logistic Regression is a benchmark to gauge the effect of feature engineering and offers a useful benchmark for low-resource scenarios.

5 Results and Discussion

We experimented with the performance of M-BERT, XLM-R, and Logistic Regression models in detecting AI-generated product reviews in Tamil and Malayalam.

M-BERT performed excellently with 0.94 precision, 0.95 recall, and 0.94 F1-scores for Tamil AI class, and 0.95, 0.94, and 0.94 respectively for HUMAN. For Malayalam, M-BERT obtained 0.91 precision for AI and 0.93 for HUMAN with respective F1-scores.

XLM-R marginally improved over M-BERT in Tamil recall (precision 0.95, recall 0.96, F1 0.92) and was considerably better for Malayalam (F1 0.75 vs. 0.68 for M-BERT), showing that it manages Malayalam's language intricacies more effectively.

Logistic Regression, with a general accuracy of 0.82, fared poorer than both M-BERT and XLM-R. Its F1-scores were 0.83 for Tamil AI, 0.81 for Tamil HUMAN, 0.83 for Malayalam AI, and 0.81 for Malayalam HUMAN, revealing the weakness of this model, particularly in dealing with Dravidian language intricacies.

Overall, both M-BERT and XLM-R performed better than Logistic Regression, with the latter being notably better for Malayalam. This indicates that transformer-based models such as M-BERT and XLM-R perform better in AI-generated review detection for Tamil and Malayalam languages, with the latter being the best fit for Malayalam.

6 Error Analysis

The mBERT model performed well in detecting AI-generated reviews in Tamil and Malayalam but faced challenges in misclassifying specific human-written reviews, especially in Tamil. This was due to linguistic features resembling those of AI-generated content. Similar misclassifications were observed in Malayalam, indicating the model's difficulty in capturing subtle contextual cues.

Despite a balanced dataset, errors, primarily false positives, highlight issues in classifying AI-generated content in low-resource languages. Fine-tuning the model's parameters is necessary to improve accuracy and make the system more reliable in classifying reviews correctly.

7 Limitations

The research is confronted with a number of limitations, mostly because of the difficulties of low-resource languages such as Tamil and Malayalam, which do not have enough annotated datasets and computational resources for successful AI-generated text detection. The poorer performance in Malayalam (macro F1-score of 0.68) as opposed to Tamil (macro F1-score of 0.90) reflects the challenge of detecting linguistic subtleties, morphology, and code-mixed forms. The use of transformer-based models like mBERT and XLM-R, although useful, is still prone to missing subtle contextual signals, and therefore misclassifies text, especially in human-composed reviews which are written in a similar AI-like style. Furthermore, the work concentrates on binary classification and does not address more complex cases, for example, mixed content of AI and human. Future research may overcome these limitations by using bigger datasets, more sophisticated fine-tuning methods, and ensemble models that combine linguistic and behavioral features for better detection performance.

8 Conclusion

Finally, a comparison of the mBERT model for detecting reviews generated using AI in Tamil and Malayalam demonstrates its advantages and limitations. The capacity for language nuance captured by it accounted for its effectiveness, obtaining F1-scores of 0.90 in Tamil and 0.68 in Malayalam, suggesting accurate performance for binary classification. Analogously, the XLM-R model also attained 0.92 for Tamil and 0.75 for Malayalam, demonstrating more competent management of language complexities.

Yet, the model was plagued with misclassifications, especially in human-like AI text, resulting in false positives. The glitches indicate shortcomings in contextual understanding, calling for enhancements. Future research should prioritize enhanced fine-tuning, data augmentation, and hybrid strategies to improve detection precision in varied linguistic contexts.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Alexis Conneau, Guillaume Lample, Sebastian Ruder, et al. 2020. Unsupervised cross-lingual representation learning. *arXiv preprint arXiv:2006.03618*.

Tommaso Fagni, Fabrizio Falchi, et al. 2021. Tweepfake: About detecting deepfake tweets. In *Proceedings of the 43rd European Conference on Information Retrieval*, pages 225–238.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, et al. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822.

Jiwei Li, Will Wang, et al. 2020. Fooling humans with ai-generated reviews: An analysis. *Journal of Artificial Intelligence Research*, 69:125–147.

Jiwei Li, Xinyuan Zhang, et al. 2022. Ai-generated reviews and their impact on online marketplaces. *Journal of Artificial Intelligence Research*, 75:1123–1145.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

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.

Lingfei Wu, Xinyuan Zhang, et al. 2021. Mind the fake review: Implications of ai-generated reviews for e-commerce. *ACM Transactions on the Web*, 15(4):1–26.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, et al. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, pages 9054–9065.

Ying Zhang et al. 2020. Overview of fake review detection methods: From heuristic rules to deep learning. *IEEE Transactions on Computational Social Systems*, 7(5):1236–1248.