KECLinguAIsts@DravidianLangTech 2025: Detecting AI-generated Product Reviews in Dravidian Languages

Malliga Subramanian[1], Rojitha R[1], Mithun Chakravarthy[1], Renusri R V[1], Kogilavani Shanmugavadivel[2]
[1]Department of CSE, Kongu Engineering College, Perundurai, Erode.
[2]Department of AI, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{rojithar.23cse, mithunchakravarthyy.23cse}@kongu.edu
renusrir.23cse@kongu.edu

Abstract

With the surge of AI-generated content in online spaces, ensuring the authenticity of product reviews has become a critical challenge. This paper addresses the task of detecting AI-generated product reviews in Dravidian languages, specifically Tamil and Malayalam, which present unique hurdles due to their complex morphology, rich syntactic structures, and code-mixed nature. We introduce a novel methodology combining machine learning classifiers with advanced multilingual transformer models to identify AI-generated reviews. Our approach not only accounts for the linguistic intricacies of these languages but also leverages domain-specific datasets to improve detection accuracy. For Tamil, we evaluate Logistic Regression, Random Forest, and XGBoost, while for Malayalam, we explore Logistic Regression, Multinomial Naive Bayes (MNB), and Support Vector Machines (SVM). Transformer-based models significantly outperform these traditional classifiers, demonstrating superior performance across multiple metrics.

1 Introduction

The rise of AI-generated content, particularly product reviews, has transformed online interactions but also raised concerns about authenticity. Detecting such reviews is especially challenging in low-resource languages like Tamil and Malayalam, which feature complex morphology, intricate syntax, and frequent code-mixing. Due to the scarcity of annotated datasets and linguistic diversity, developing robust detection models for these languages remains difficult. This study explores using traditional machine learning classifiers—Logistic Regression, Random Forest, and XGBoost for Tamil, and Logistic Regression, Multinomial Naive Bayes (MNB), and SVM for Malayalam—to detect AI-generated product reviews, demonstrating their effectiveness in addressing the unique challenges of low-resource languages.Chakravarthi et al., 2021

2 Literature Survey

Research on AI-generated product review detection has mainly focused on high-resource languages like English, with limited attention to low-resource Dravidian languages like Tamil and Malayalam. These languages pose challenges due to linguistic complexities such as code-mixing and context-dependent nuances. While some multilingual datasets exist, research specifically addressing AI-generated content in product reviews remains scarce.Baiju, 2023 The lack of large annotated datasets and the subtle nature of AI-generated reviews further complicate detection. This survey reviews models from DravidianLangTech@NAACL 2025 for AI-generated review detection. Advancements in transfer learning and multilingual embeddings could help address these challenges.Premjith et al., 2025

2.1 Detection of AI-Generated Content in English and Major Languages

Early methods for detecting AI-generated content relied on handcrafted features like text perplexity, word-level n-grams, and syntactic patterns. While effective for earlier AI systems, they struggled with modern generative models. Recent advancements, particularly transformer-based models like RoBERTa, GPT detectors, and BERT, enhance detection by analyzing contextual relationships and embedding patterns. These models identify subtle inconsistencies, such as repetitive phrasing and unnatural expressions, indicative of AI-generated text. However, their effectiveness is limited in Dravidian languages due to a lack of pre-trained models, emphasizing the need for fine-tuning on language-specific data.Muneer and Basheer, 2023

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 274–277

2.2 Detection of AI-Generated Reviews in Dravidian Languages

Detecting AI-generated product reviews in Tamil and Malayalam is an underexplored area, crucial due to the rising prevalence of such content online. Dravidian languages present challenges like complex syntax, rich morphology, and frequent code-mixing, complicating detection. The scarcity of annotated datasets for AI-generated content further limits model development. Current research mainly focuses on traditional text classification using Logistic Regression, Random Forest, and SVM, but these models struggle with subtle AI markers in code-mixed or informal text.Gautam and Bharathi, 2021 Transformer-based models like BERT and XLM-Roberta show promise but require extensive fine-tuning and dataset augmentation to address linguistic diversity.

2.3 Deep Learning and Transformers

Advancements in deep learning, particularly with transformer-based models like BERT, GPT, and XLM-R, have shown potential in addressing the challenges of detecting AI-generated reviews. These models use self-attention mechanisms to capture complex word relationships, enabling the detection of subtle AI-generated patterns. Unlike traditional methods relying on feature extraction, transformers analyze context holistically, making them effective for identifying generative AI outputs. However, their performance in Tamil and Malayalam is limited by the lack of large-scale annotated datasets.Sebastian, 2023 Fine-tuning them on Dravidian language-specific corpora is essential for better detection, especially in code-mixed and context-dependent reviews. Future improvements include using larger annotated datasets and transformer-based models like BERT or XLM-R to enhance contextual understanding and detection accuracy.

2.4 Practical Constraints of Transformer Models

Transformer models like BERT, mBERT, and XLM-R require significant computational power, making real-time deployment challenging. Their training time is much longer than traditional models, especially with multiple epochs. Large annotated datasets are essential, which are often scarce for low-resource languages like Tamil and Malayalam. Fine-tuning these models is complex, requiring careful hyperparameter tuning. Unlike traditional models, transformers lack interpretability, making their predictions harder to explain.

3 Materials and Methods

3.1 Taskset Description

This study focuses on detecting AI-generated product reviews in Tamil and Malayalam. The dataset consists of 1,608 training samples (808 Tamil, 800 Malayalam) and 300 test samples (100 Tamil, 200 Malayalam), each labeled as "AI" or "HUMAN", with a sample format shown in Figures 1 and 2. The task involves identifying whether a given review is AI-generated or human-written. Six machine learning models—Logistic Regression, Random Forest, XGBoost, Support Vector Machines (SVM), and Multinomial Naive Bayes (MNB)—are evaluated based on accuracy, precision, recall, and F1-score. The study aims to develop a robust framework for detecting AI-generated content in low-resource languages like Tamil and Malayalam.

ID	DATA	LABEL
TAM_HUAI_TR_386	இந்த கேபேஸ் வாஷ் சருமத்தை உலர்த்துகிறது.	AI
TAM_HUAI_TR_396	இந்த ஜ கிரீம் எரிச்சலை ஏற்படுத்துகிறது.	AI
TAM_HUAI_TR_401	இந்த நெயில் கபைல் விரலை கீறுகிறது.	AI
TAM_HUAI_TR_420	அதிகமா செலவு பண்ண வேண்டி இருக்கு	HUMAN
TAM_HUAI_TR_504	எனக்கு ஏற்ற அளவு செருப்பு கிடைக்கவில்லை	HUMAN

Figure 1: Sample training texts from Tamil dataset

ID	DATA	LABEL
MAL_HUAI_TR_012	കാജൽ റിമൂവറിന് ഈ പ്രോഡക്റ്റ് നല്ലതാണ്	HUMAN
MAL_HUAI_TR_021	നമുക്ക് ആവശ്യമായ സാധനങ്ങൾ അടിപൊളി	HUMAN
MAL_HUAI_TR_153	8 ആഴ്ച യൂസ് ചെയ്താൽ മതി. പെർമനൻറ് ആണ്	HUMAN
MAL_HUAI_TR_762	മീൻ കറിയിനോടൊപ്പം പപ്പടവും, പുളിശ്ശേരിയും പൊളിയാ!	AI
MAL_HUAI_TR_766	ഇവിടെ എവിടെ എങ്കിലും നാടൻ ഭക്ഷണവും ലഭിക്കും.	AI

Figure 2: Sample training texts from Malayalam dataset

3.2 Preprocessing and Feature Extraction

Preprocessing transformed the Tamil and Malayalam product review data into a machine learning-friendly format. The dataset, labeled as "AI" or "HUMAN", was cleaned by handling missing values, standardizing text, and converting it to lowercase. Stop words were removed, and TF-IDF vectorization was applied to extract numerical features. Tokenization and stemming were used to break text into words and reduce them to base forms. These steps ensured high-quality input for effective detection of AI-generated reviews.Sinthusha et al., 2025

3.3 Models and Methodology

This study uses machine learning models to detect AI-generated product reviews in Tamil and Malayalam text. For Tamil, we employed Logistic Regression (LR), Random Forest (RF), and XGBoost. LR was chosen for its interpretability, RF for its ensemble learning capabilities, and XGBoost for its ability to capture complex patterns. For Malayalam, we used Support Vector Machine (SVM), Multinomial Naive Bayes (MNB), and Logistic Regression LR. SVM handles high-dimensional data, MNB is effective for word frequency distributions, and LR provides consistent performance.Priyadharshini et al., 2021 Text data was preprocessed using the TF-IDF vectorization technique, converting text into numerical features while handling missing values and code-mixing. The models were trained and evaluated based on accuracy, precision, recall, and F1-score to assess their effectiveness in detecting AI-generated reviews.

4 Results and Discussion

The study on detecting AI-generated product reviews in Tamil and Malayalam demonstrated that while various machine learning models, including Logistic Regression (LR), Random Forest (RF), and XGBoost, performed well, Logistic Regression outperformed the others in both languages. This makes Logistic Regression the most effective model for detecting AI-generated content in Tamil and Malayalam, showing its robustness in handling the intricacies of these low-resource languages. Figure 3 illustrates the Confusion Matrix for the high-performing model (LR) in Tamil and Malayalam.

4.1 Performance Metrics

The models were evaluated using Accuracy, Precision, Recall, and F1-Score. Accuracy measures the proportion of correctly classified reviews, while Precision indicates the percentage of correctly identified AI-generated reviews. Recall reflects the proportion of actual AI reviews detected, and F1-Score balances Precision and Recall, which is important for imbalanced datasets. These metrics are vital for assessing model performance. Table 1 shows the results on the training dataset, while Table 2 presents the test dataset performance, highlighting the variation in model effectiveness across datasets.

Classifiers	Class Labels	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression (Tam)	AI	86	85	89	87
	HUMAN	86	88	84	86
Random Forest (Tam)	AI	91	88	94	91
	HUMAN	91	93	88	90
XGBoost (Tam)	AI	84	86	79	82
	HUMAN	84	83	88	85
Logistic Regression (Mal)	AI	74	77	70	73
	HUMAN	74	72	79	75
MNB (Mal)	AI	86	85	89	87
	HUMAN	86	88	84	86
SVM (Mal)	AI	79	78	80	79
	HUMAN	79	79	78	78

Table 1: Performance of Classifiers for AI and HUMAN Text Detection in Tamil and Malayalam(Training Dataset)

Classifiers	Class Labels	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression (Tam)	AI	68	65	71	68
	HUMAN	68	71	65	68
Random Forest (Tam)	AI	68	67	67	67
	HUMAN	68	69	69	69
XGBoost (Tam)	AI	68	70	58	64
	HUMAN	68	67	77	71
Logistic Regression (Mal)	AI	67	68	64	66
	HUMAN	67	66	70	68
MNB (Mal)	AI	65	76	45	57
	HUMAN	65	61	86	71
SVM (Mal)	AI	65	67	62	64
	HUMAN	65	64	69	67

Table 2: Performance of Classifiers for AI and HUMAN Text Detection in Tamil and Malayalam(Test Dataset)

4.2 Error Analysis

Despite Logistic Regression achieving the highest accuracy, it misclassified several instances, predicting AI-generated reviews as HUMAN and vice versa. Figure 4 illustrates a few such examples, likely due to TF-IDF's inability to capture contextual nuances and challenges with code-mixed text.

ID	DATA	Predictions	Real_Predictions
TAM_HUAI_TE_086	பேக்கிங் சரியில்லை	AI	HUMAN
TAM_HUAI_TE_088	மிகவும் வாசனை உள்ள பொருள்	AI	HUMAN
MAL_HUAI_TE_104	ചുടുള്ള ഭക്ഷണം അലുമിനിയം ഫോയിലിൽ	HUMAN	AI
MAL_HUAI_TE_138	ഇക്കാലത്ത് ഒക്കെ ടാറ്റയുടെ ഡിസൈൻ കണ്ട്	HUMAN	AI

Figure 4: Example of a misclassified Tamil-Malayalam code-mixed text by LR model.

5 Limitations

Detecting AI-generated reviews in Tamil and Malayalam is challenging due to the lack of annotated datasets, limiting model performance. The complex syntax, rich morphology, and frequent code-mixing make feature extraction difficult. Traditional models relying on basic features like n-grams fail to capture subtle AI-generated patterns. Additionally, distinguishing human-like fluency in AI-generated content from human-authored text requires deeper linguistic understanding.Kumaresan and Pal, 2021 Cultural and contextual variations further complicate detection without robust language-specific datasets and tools.

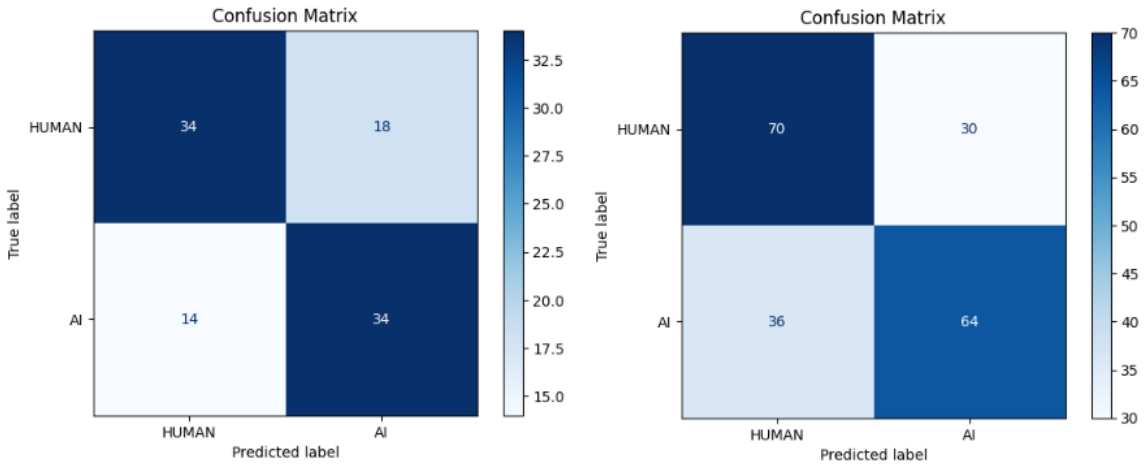

Figure 3: Confusion Matrix for high performing model(LR) in Tamil and Malayalam

6 Conclusion

This study explored detecting AI-generated and human-written reviews in Tamil and Malayalam using machine learning. Logistic Regression emerged as the most effective model, handling challenges in low-resource, morphologically rich languages.Nair et al., 2014 The results highlight the importance of tailored preprocessing and feature extraction techniques. Despite the promising outcomes, Future work should focus on creating larger annotated datasets and incorporating transformer-based models for better contextual understanding.Zhu and Dong, 2020 Hybrid approaches combining traditional and advanced models can further enhance detection accuracy in diverse linguistic contexts. The datasets and implementation code utilized in this research are publicly available at GitHub Repository to support reproducibility and further research.

References

K. B. Baiju. 2023. *Pattern primitive based malayalam handwritten character recognition studies for real-time applications*. Ph.D. thesis, Department of Computer Science, University of Calicut.

Bharathi Raja Chakravarthi et al. 2021. Dravidian-multimodality: A dataset for multi-modal sentiment analysis in tamil and malayalam. *arXiv preprint arXiv:2106.04853*.

Abhishek Kumar Gautam and B. Bharathi. 2021. Rnn's vs transformers: Training language models on deficit datasets. In *FIRE (Working Notes)*, pages 737–743.

Kumar Kumaresan and Kingston Pal. 2021. Dravidian-multimodality: A dataset for multi-modal sentiment analysis in tamil and malayalam. *arXiv preprint arXiv:2106.04853*.

V. K. Muneer and K. P. Mohamed Basheer. 2023. A collaborative destination recommender model in dravidian language by social media analysis. In *Proceedings of Data Analytics and Management: ICDAM 2022*, pages 541–551. Springer Nature Singapore.

Deepu S. Nair et al. 2014. Sentima-sentiment extraction for malayalam. In *2014 International conference on advances in computing, communications and informatics (ICACCI)*, pages 1719–1723. IEEE.

B Premjith, Nandhini K, Bharathi Raja Chakravarthi, Thenmozhi Durairaj, Balasubramanian Palani, Sajeetha Thavareesan, and Prasanna Kumar Kumaresan. 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.

Ruba Priyadharshini et al. 2021. Overview of the dravidiancodemix 2021 shared task on sentiment detection in tamil, malayalam, and kannada. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 4–6.

Mary Priya Sebastian. 2023. Malayalam natural language processing: challenges in building a phrase-based statistical machine translation system. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–51.

AV Ann Sinthusha, E. Y. A. Charles, and Ruvan Weerasinghe. 2025. Machine reading comprehension for the tamil language with translated squad. *IEEE Access*.

Yueying Zhu and Kunjie Dong. 2020. Yun111@ dravidian-codemix-fire2020: Sentiment analysis of dravidian code mixed text. In *FIRE (Working Notes)*, pages 628–634.