TeamVision@DravidianLangTech 2025: Detecting AI generated product reviews in Dravidian Languages

Shankari S R, Sarumathi P, Bharathi B
Department of Computer Science and Engineering
Sri Sivasubramania Nadar College of Engineering
shankari2210607@ssn.edu.in
sarumathi2210526@ssn.edu.in
bharathib@ssn.edu.in

Abstract

Recent advancements in natural language processing (NLP) have enabled artificial intelligence (AI) models to generate product reviews that are indistinguishable from those written by humans. To address these concerns, this study proposes an effective AI detector model capable of differentiating between AI-generated and human-written product reviews. Our methodology incorporates various machine learning techniques, including Naive Bayes, Random Forest, Logistic Regression, SVM, and deep learning approaches based on the BERT architecture. Our findings reveal that BERT outperforms other models in detecting AI-generated content in both Tamil product reviews and Malayalam product reviews.

1 Introduction

Online product reviews play a vital role in shaping consumer behavior and market dynamics. However, the rise of AI-generated reviews poses a threat to the reliability of online platforms by enabling misleading content. Detecting such reviews is crucial to maintaining consumer trust and informed decision-making. This challenge is amplified in low-resource languages like Malayalam and Tamil, which feature complex linguistic structures and limited annotated datasets. This research focuses on developing models to identify AI-generated reviews in these languages, leveraging both machine learning and deep learning techniques. By addressing this gap, the study contributes to AI content detection, supports linguistic diversity, and enhances trust in digital ecosystems.

[1].

[1] https://github.com/Shankarisr/TeamVision-Detecting-AI-generated-product-reviews.git

2 Related Work

Natural Language Processing (NLP) and machine learning have been widely used to detect AI-generated text. Prova (2023) explored various NLP and machine learning-based approaches to identify synthetic text, emphasizing their effectiveness in distinguishing AI-generated content from human-written text (Prova, 2023).

(Akram, 2023) addressed the growing need for reliable evaluation by developing a multi-domain dataset designed to test state-of-the-art APIs and tools for identifying AI-generated content. Building upon this foundation, our study investigates the effectiveness of AI text detection methods.

Desaire et al. (2023) findings revealed that domain-specific prompts could influence the detectability of AI-generated content, making it more challenging for existing detection models to distinguish between human and synthetic text (Desaire et al., 2023).

Gritsay et al. (2022) examined the effectiveness of AI text detection and emphasized the need for more tokens to improve accuracy (Gritsay et al., 2022). (Shimi et al., 2024) addressed an empirical analysis of language detection in Dravidian languages, focusing on challenges and advancements specific to languages like Tamil, Malayalam, Kannada, and Telugu.

H. B. S. and Rangan (2020) conducted a comprehensive survey on Indian regional language processing, highlighting the challenges and advancements in NLP for languages like Tamil and Malayalam (S. and Rangan, 2020).

Ponnusamy (2023) explored the use of ChatGPT-3 models for Tamil text generation, focusing on how AI models can be leveraged to generate coherent and contextually relevant text in Tamil (Ponnusamy, 2023).

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 694–698

3 Dataset

The goal of this task is to develop a model that can effectively detect AI-generated product reviews in Dravidian languages like Tamil and Malayalam. The dataset used for this purpose is sourced from the Detecting AI-generated product reviews in Dravidian languages, provided by DravidianLangTech@NAACL 2025 (Premjith et al., 2025). The training dataset includes the fields id, data, and label, supporting a supervised learning approach. On the other hand, the testing dataset contains only the id and data, which are used exclusively for making predictions. The dataset descriptions are given in Table 1.

Category	Tamil		Malayalam	
	Train	Test	Train	Test
AI	405	-	400	-
Human	403	-	400	-
Total	808	200	800	100

Table 1: Summary of the training and testing dataset entries for Tamil and Malayalam.

4 Methodology

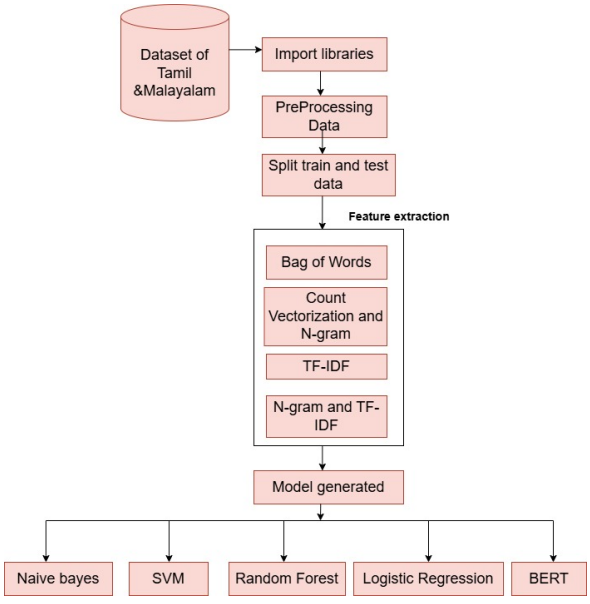

Figure 1: The proposed methodology of the work

4.1 Preprocessing

Text preprocessing is a crucial step in NLP that prepares raw text data for analysis by cleaning and structuring it. It involves several techniques to reduce noise, standardize data, and focus on meaningful content. Stopword Removal filters out common words that carry little semantic meaning, helping to reduce noise and emphasize significant terms. Removing Unwanted characters ensures retaining only relevant content. Tokenization breaks text into smaller units for analysis. Stemming reduces words to their base form by stripping suffixes, simplifying text processing. Lemmatization provides linguistically accurate base forms by grouping words with their root words. This preprocessing pipeline ensures that data is clean, standardized, and optimized for downstream tasks, improving computational efficiency and model accuracy.Figure 2 shows examples of techniques used to reduce noise.

Technique	Tamil Example	Malayalam Example
Stopword Removal	Original: இந்த போன் உண்மையில் மிகவும் நல்லது மற்றும் பயனுள்ளதாக உள்ளது. After: போன் நல்லது பயனுள்ளது.	Original: ഈ ഫോൺ വളരെ നല്ലതാണ്, ഉപകാരപ്രദവുമാണ്. After: ഫോൺ നല്ലത് ഉപകാരപ്രദം.
Removing Unwanted Characters	Original: அற்புதமான தயாரிப்பு!!! பாருங்கள்: www.example.com. After: அற்புதமான தயாரிப்பு பாருங்கள்.	Original: അസാധാരണമായ ഉൽപ്പന്നം!!! പരിശോധിക്കുക: www.example.com. After: അസാധാരണമായ ഉൽപ്പന്നം പരിശോധിക്കുക.
Tokenization	Original: இந்த போன் மிகவும் நல்லது. After: இந்த போன், மிகவும், நல்லது.	Original: ഈ ഫോൺ വളരെ നല്ലതാണ്. After: ഈ, ഫോൺ, വളരെ, നല്ലതാണ്.
Stemming	Original: இருப்பது (being) After:"இரு	Original: ഇരിക്കുന്നു (is) After: ഇരു
Lemmatization	Original: சிறந்த (better) After: சிறந்த	Original: മികച്ച (better) After: മികച്ച

Figure 2: Examples of preprocess techniques

4.2 Feature Engineering

In NLP, feature extraction techniques like BoW, TF-IDF, and n-grams convert text into numerical formats for machine learning. BoW counts word frequency, while TF-IDF weights words based on importance. Using TF-IDF or Count Vectorization with n-grams captures word context. Performance is assessed using F1-Score and accuracy. Figure 1 illustrates the architecture.

4.3 Model Generate

4.3.1 Naïve Bayes

Naïve Bayes, a probabilistic classifier based on Bayes' theorem, assumes feature independence. It achieved 90.12% accuracy (F1: 0.91 AI, 0.88 human) in Tamil and 76.87% accuracy (F1: 0.77 AI, 0.77 human) in Malayalam.

4.3.2 Logistic Regression

Logistic Regression predicts categorical outcomes using probability modeling. It achieved 88.27% accuracy (F1: 0.89 AI, 0.89 human) in Tamil and 76.87% accuracy (F1: 0.75 AI, 0.75 human) in Malayalam.

4.3.3 Support Vector Machine (SVM)

SVM finds the optimal hyperplane for classification. It achieved 89.5% accuracy (F1: 0.89 AI, 0.89 human) in Tamil and 75.62% accuracy (F1: 0.75 AI, 0.76 human) in Malayalam.

4.3.4 Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to improve classification accuracy and reduce overfitting. Achieved 85.8% accuracy (F1: 0.86 AI, 0.91 human) in Tamil and 78.75% accuracy (F1: 0.8 AI, 0.8 human) in Malayalam.

4.3.5 KNN

KNN classifies data based on the majority of nearest neighbors. It achieved 85.8% accuracy (F1: 0.86 for AI human text) in Tamil and 73.12% accuracy (F1: 0.74 for AI, 0.72 for human text) in Malayalam.

4.3.6 BERT

The proposed model uses Multilingual BERT (mBERT) because it supports multiple languages, including Tamil and Malayalam, without requiring separate models for each language.Achieved an accuracy of 90%, with F1 scores of 0.9 for AI and 0.89 for human text.Achieved the highest accuracy of 95%, with F1 scores of 1 for both AI and human text in Malayalam.

4.3.7 Justification

The BERT-based model excelled due to its deep contextual understanding, while Naïve Bayes and Logistic Regression struggled with complex patterns. SVM performed well but was computationally expensive, and Random Forest lacked contextual depth, affecting AI text detection.

5 Results and Discussion

Model	Feature	Prec.	Rec.	F1
NB	BoW	0.88	0.87	0.88
	TF-IDF	0.88	0.87	0.87
	ngram+tf-idf	0.91	0.90	0.89
	CVec + ngram	0.92	0.90	0.90
LR	BoW	0.87	0.87	0.87
	tf-idf	0.87	0.87	0.87
	ngram+tf-idf	0.88	0.88	0.88
	CVec + ngram	0.88	0.88	0.88
SVM	BoW	0.89	0.89	0.89
	ngram+tf-idf	0.90	0.90	0.90
	CVec+ngram	0.88	0.88	0.88
	tf-idf	0.90	0.89	0.89
RF	BoW	0.85	0.85	0.85
	tf-idf	0.86	0.86	0.86
	CVec+ngram	0.84	0.83	0.83
	ngram+tf-idf	0.86	0.86	0.86
DT	BoW	0.83	0.83	0.83
	tf-idf	0.81	0.81	0.81
	ngram + tf-idf	0.82	0.82	0.82
	CVec + ngram	0.85	0.85	0.85
KNN	BoW	0.78	0.63	0.63
	ngram + tf-idf	0.71	0.46	0.44
	CVec + ngram	0.86	0.36	0.36
BERT	**BERT Emb.**	**0.98**	**0.98**	**0.98**

Table 2: Performance Comparison of Models on Tamil Dataset (Premjith et al., 2025)

The performance evaluation of various classifiers on the Tamil and Malayalam shown in Table 2 and Table 3 highlights significant differences in effectiveness across models. BERT emerges as the most accurate classifier, achieving the highest precision, recall, and F1 scores for both AI and human text classification. Specifically, for the Malayalam, BERT reaches an impressive F1 score of 96 for both AI and human text, while in the Tamil, it achieves 97 for AI and 99 for human text. These results emphasize the power of deep learning-based transformer models in understanding complex linguistic patterns, even in low-resource languages. These results suggest that BERT is the go-to choice for classifying AI-human text in languages like Tamil and Malayalam. This highlights a key di-

Model	Feature	Prec.	Rec.	F1
NB	BoW	0.76	0.76	0.76
	tf-idf	0.75	0.75	0.75
	ngram+tf-idf	0.76	0.76	0.76
	CVec+ngram	0.77	0.77	0.77
LR	BoW	0.77	0.76	0.76
	tf-idf	0.76	0.76	0.76
	ngram+tf-idf	0.77	0.77	0.77
	CVec+ngram	0.77	0.77	0.77
SVM	BoW	0.70	0.69	0.69
	tf-idf	0.73	0.73	0.73
	CVec+ngram	0.72	0.72	0.72
	N-gram+tf-idf	0.76	0.76	0.76
RF	BoW	0.78	0.75	0.75
	tf-idf	0.79	0.79	0.79
	CVec+ngram	0.76	0.74	0.74
	N-gram+tf-idf	0.75	0.75	0.75
DT	BoW	0.73	0.72	0.72
	tf-idf	0.69	0.69	0.69
	CVec+ngram	0.74	0.74	0.74
KNN	BoW	0.59	0.52	0.52
	tf-idf	0.74	0.73	0.73
	ngram+tf-idf	0.71	0.71	0.71
	CVec+ngram	0.66	0.39	0.39
BERT	**BERT Emb.**	**0.96**	**0.96**	**0.96**

Table 3: Performance Comparison of Models on Malayalam Dataset (Premjith et al., 2025)

rection for future research: fine-tuning transformer models to better handle low-resource languages, using their strong ability to understand context and generalize across data to boost classification accuracy even further.

5.1 Comparison with Existing AI Text Detection Tools

The proposed BERT-based model surpasses existing AI text detection tools in handling Tamil and Malayalam, while tools like GPTZero, OpenAI AI Text Classifier, GLTR, and Turnitin primarily focus on English. Unlike statistical or probabilistic models, the proposed approach leverages TF-IDF, n-grams, and contextual embeddings, allowing better customization and adaptability for low-resource languages. While existing tools lack fine-tuning for non-English texts, the proposed model effectively detects AI-generated reviews with moderate explainability, making it more suitable for review detection in underrepresented languages compared to general-purpose detection tools.

6 Conclusions

The experimental results on Malayalam and Tamil datasets demonstrate that transformer-based models, particularly BERT, significantly outperform traditional machine learning approaches in classification accuracy. BERT achieves the highest precision, recall, and F1 scores across both datasets, reinforcing its effectiveness in handling complex linguistic structures in low-resource languages. While traditional classifiers like Naïve Bayes, Logistic Regression, SVM, and Random Forest show moderate performance, models like Decision Tree and KNN struggle to generalize effectively. These findings highlight the importance of leveraging deep learning models for AI-human text classification, ensuring reliable detection methods in the face of rapidly advancing AI-generated content.

7 Limitations

Tamil and Malayalam lack large, high-quality datasets, making AI models prone to bias and inaccuracies.Many users blend Tamil/Malayalam with English or use Romanized script, which traditional models struggle to process. Rich morphology in these languages makes tokenization and feature extraction difficult, reducing model accuracy. AI models, including BERT, often misinterpret sarcasm and subtle sentiments, leading to errors.Training BERT for Tamil and Malayalam requires significant resources, limiting practical use.Language and user reviews evolve over time, causing model degradation. Continuous updates and retraining are essential to maintain classifier accuracy in real-world applications.

8 Error Analysis

False Positives: Formal or repetitive genuine reviews misclassified. False Negatives: AI-generated reviews mimicking humans went undetected. Language Issues: Struggled with code-mixed text and dialects. Improvements: Train on diverse data, refine code-mixed handling, and add context-aware features.

References

Arslan Akram. 2023. An empirical study of ai generated text detection tools. *ArXiv*, abs/2310.01423.

Heather Desaire, Andrea E. Chua, Min Gyu Kim, and David Hua. 2023. Accurately detecting ai text when chatgpt is told to write like a chemist. *Cell Reports Physical Science*, 4(11):101672.

German Gritsay, Andrey Grabovoy, and Yu. V. Chekhovich. 2022. Automatic detection of machine generated texts: Need more tokens. *2022 Ivannikov Memorial Workshop (IVMEM)*, pages 20–26.

R. Ponnusamy. 2023. Tamil text generation using chatgpt-3 models. [Online]. Available: ponnusamyr@citchennai.net.

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.

N. N. I. Prova. 2023. Detecting ai generated text based on nlp and machine learning approaches. [Online]. Available: nuzhatnsu@gmail.com.

H. B. S. and R. K. Rangan. 2020. A comprehensive survey on indian regional language processing. *SN Applied Sciences*, 2(7):1–16.

G Shimi, CJ Mahibha, and D Thenmozhi. 2024. An empirical analysis of language detection in dravidian languages. *Indian Journal of Science and Technology*, 17(15):1515–1526.