Team-Risers@DravidianLangTech 2025: AI-Generated Product Review Detection in Dravidian Languages Using Transformer-Based Embeddings

Sai Sathvik P[1], Muralidhar Palli[1], Keerthana NNL[1], Balasubramanian Palani[1], Jobin Jose[1], Siranjeevi Rajamanickam[2]

Department of Computer Science and Engineering, IIIT Kottayam, Kerala, India[1]
Dept of Computer Engineering, Govt. Polytechnic College, Trichy, India.[2]
(psaisathvik612,muralidharpalli12345,nnl.Keerthana)@gmail.com,
pbala@iiitkottayam.ac.in, jobin@iiitkottayam.ac.in, rajasiranjeevi@gmail.com

Abstract

Online product reviews influence customer choices and company reputations. However, companies can counter negative reviews by generating fake reviews that portray their products positively. These fake reviews lead to legal disputes and concerns, particularly because AI detection tools are limited in low-resource languages such as Tamil and Malayalam. To address this, we use machine learning and deep learning techniques to identify AI-generated reviews. We utilize Tamil BERT and Malayalam BERT in the embedding layer to extract contextual features. These features are sent to a Feedforward Neural Network (FFN) with softmax to classify reviews as AI-generated or not. The performance of the model is evaluated on the dataset. The results show that the transformer-based embedding achieves a better accuracy of 95.68% on Tamil data and an accuracy of 88.75% on Malayalam data.

1 Introduction

In today's digital era, online reviews influence purchasing decisions. Customers rely heavily on reviews when deciding which products to buy on e-commerce sites. However, with AI advancements, companies started leveraging AI to enhance brand credibility and increase awareness by posting fake reviews, making it difficult for users to separate fact from fiction. It's hard to distinguish between real and fake reviews, which spreads false information. Detecting low-resource languages, such as Dravidian languages, lags behind more commonly used languages like English and Spanish. This study aims to detect AI-produced reviews in Dravidian languages, mainly Tamil and Malayalam. Using advanced NLP techniques and pre-trained language models, our objective is to improve the trustworthiness of online reviews and have healthy competition within the digital marketplace.

2 Related Work

The area of concern is the evaluation and detection of AI-generated reviews in languages with fewer resources, such as Tamil and Malayalam, which shows various challenges due to their intricate morphology and complicated syntactic structures.

Recent studies have explored the application of machine learning and transfer learning models to detect AI-generated reviews. (Kumar et al., 2024) used models of token and paraphrase style review generation with Term Frequency to prove their effectiveness. (Al-Adhaileh and Alsaade, 2022) called attention to the capabilities of Bidirectional Long Short Term Memory (BiLSTM) networks that outperformed the CNN in the fake review detection in low-resource languages. A study by (Abdedaiem et al., 2023) highlighted a few-shot learning approach through sentence transformers to detect fake news in Algerian Arabic, indicating that a similar approach could be used for certain Dravidian languages.

In the context of Dravidian languages, research has predominantly focused on fake news detection, hate speech classification, and sentiment analysis. (Raja et al., 2023) proposed a transfer learning-based approach with adaptive fine-tuning for detecting fake news in Tamil and Malayalam, showing that domain-adaptive fine-tuning improves performance, (Roy et al., 2022) introduced a deep ensemble framework for hate speech and offensive language detection, emphasizing the necessity of language-specific models.(Mandalam and Sharma, 2021) explored sub-word representations, word embeddings, and hybrid models for Tamil-English and Malayalam-English sentiment classification, highlighting the impact of preprocessing and feature engineering.

Despite these advancements, research on AI-generated product reviews in Dravidian languages is still lacking. This study builds upon existing

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 153–157
May 3, 2025 ©2025 Association for Computational Linguistics

work in fake news detection, hate speech classification, and sentiment analysis by leveraging Tamil-BERT and Malayalam-BERT along with advanced fine-tuning techniques. By adopting state-of-the-art transformer-based models and optimizing preprocessing strategies, this research aims to bridge the gap in AI-generated reviews detection for these languages.

3 Proposed Methodology

Figure 1 demonstrates the workflow of the proposed architecture to determine if product reviews can be identified in Tamil and Malayalam languages by taking advantage of transformer architectures.

3.1 Text Preprocessing

In the NLP model, preprocessing steps are crucial for cleansing and standardizing input data for pre-trained models. Initially, these steps involved removing noise to focus on linguistic content. The text was then segmented into sentences and tokenized using language-specific tokenizers from Tamil-BERT and Malayalam-BERT.

Next, WordPiece tokenization was applied, effectively handling morphologically rich languages like Tamil and Malayalam by decomposing infrequent or compound words into subwords, preserving semantic relationships. Finally, dynamic sequence length normalization was implemented using Hugging-Face's DataCollatorWithPadding, applying uniform padding to each input sequence in a batch for compatibility with the transformer architecture while enhancing training efficiency.

3.2 Embedding Layer

Embeddings are numerical representations of textual data that transform words or phrases into dense, continuous vector spaces. This transformation allows text to be processed by machine learning and deep learning models.

3.2.1 Classical Text Encoding

Classical text encoding methods transform textual data into numerical representations for machine learning models. Two widely used approaches are Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF).

BoW: This model represents text as a collection of word occurrences without considering word order or context. Given a corpus, it constructs a vocabulary and represents each document as a vector of word frequencies. Formally, for a document d in a corpus D, the BoW representation is given by Eq.(1):

$$\text{BoW}(t, d) = \text{Count}(t, d) \tag{1}$$

where, $\text{Count}(t, d)$ is the number of times term t appears in document d. This method provides a simple and efficient representation but lacks semantic understanding.

TF-IDF: This improves upon BoW by weighting terms based on their importance within the corpus. The TF-IDF score for a term t in a document d is given by Eq.(2):

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \cdot \text{IDF}(t) \tag{2}$$

Where:

- $\text{TF}(t, d) = \frac{\text{Count}(t,d)}{\text{Total terms in } d}$ represents term frequency, and
- $\text{IDF}(t) = \log\left(\frac{|D|}{1+|\{d \in D : t \in d\}|}\right)$ accounts for how commonly a term appears across documents, reducing the weight of frequently occurring words.

While BoW captures raw word counts, TF-IDF enhances representation by emphasizing important terms, making it more effective for tasks like text classification and retrieval.

3.2.2 Transformer-Based Embedding

The transformer-based approach utilizes Tamil-BERT and Malayalam-BERT to generate dense, contextual embeddings through a multi-head self-attention mechanism.

Tokenization: Input text is tokenized into subwords using language-specific tokenizers. For a sequence $X = [x_1, x_2, \ldots, x_n]$, tokens are embedded as in Eq.(3) as follows:

$$e_i = W_e \cdot x_i + p_i \tag{3}$$

where W_e is the embedding matrix, and p_i is the positional embedding.

Self-Attention: Relationships between tokens are modelled using self-attention as shown in Eq.(4):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V \tag{4}$$

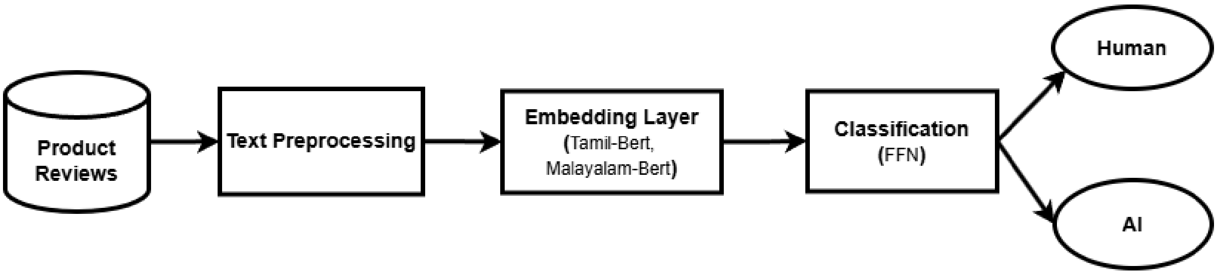


Figure 1: Proposed architecture for AI-generated review detection

where Q, K, V are derived from the input embeddings, and d_k is the key dimension. This mechanism enables the model to capture long-range dependencies, which is essential for context-rich languages.

3.3 Classification

To categorise reviews into human or AI-written, a Feedforward Neural Network (FFN) is employed, which has pre-trained contextual embeddings. The network runs the embeddings through several hidden layers by applying GeLU activation for the embedding 'hidden' layers, whereas the output layer is trained with Softmax to generate class probabilities. The class with a higher probability is predicted as 1 for Human and 0 for AI.

4 Experiment

This section provides an extensive overview of the experimental setup used for training and evaluation and the reference data sets used in this research.

4.1 Experiment Setup

The testing was effectively carried out on Google Colab, leveraging its resources to fine-tune transformer neural network models. Colab proved to be an essential platform, meeting the strict demands of these models. The dataset was split into training and testing sets in an 80:20 ratio, ensuring each set included a balanced mix of real and AI-generated reviews. The training process employed the Hugging Face Trainer API, which streamlined the automation of gradient computations, optimizations, and evaluations, making the training highly efficient.

4.2 Dataset

The dataset used in this study was sourced from the shared task (Premjith et al., 2025). Table 1 summarizes the datasets utilized for detecting AI-generated reviews.

Table 1: Summary of datasets

	Reviews	Count
Tamil	Human	403
	AI	405
Malayalam	Human	400
	AI	400

Word Distribution The dataset reveals differences in review lengths between AI-generated and human-written reviews. As illustrated in Figures 2a and 2b, AI-generated reviews are generally shorter and more concentrated around a lower word count, while human-written reviews display a broader distribution with longer text samples. The Tamil dataset peaks around 10–15 words for AI-generated reviews, whereas human-written reviews encompass a wider range, often exceeding 20 words. Similarly, the Malayalam dataset exhibits a similar pattern, with AI-generated reviews clustering around shorter lengths, while human reviews demonstrate greater variability in length.

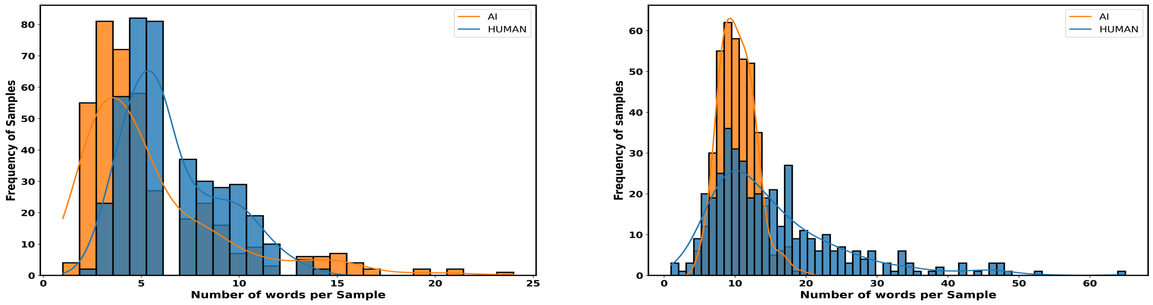
4.3 Evaluation Metrics

Standard metrics were employed to assess the performance of the classification model: Accuracy, Precision, Recall, F1-Score and Macro F1-Score. These metrics offer a clear perspective on the model's capability to distinguish between real and AI-generated reviews.

$$\text{Accuracy} = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|} \quad (5)$$

where |TP| = Count of true positive reviews, |FP| = Count of false positive reviews, |FN| = Count of false negative reviews, |TN| = Count of true negative reviews.

$$\text{Precision} = \frac{|TP|}{|TP| + |FP|} \quad (6)$$



(a) Tamil dataset

(b) Malayalam dataset

Figure 2: Word Distribution on AI vs Human

$$\text{Recall} = \frac{|TP|}{|TP| + |FN|} \tag{7}$$

$$\text{F1-score} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{8}$$

$$\text{Macro F1-score} = \frac{1}{N}\sum_{i=1}^{N} \text{F1-score}_i \tag{9}$$

where, N is the number of classes, and F1-score$_i$ is the F1-score for class i.

5 Results

The performance of various machine learning (ML) models was evaluated on the test datasets for Tamil and Malayalam reviews. Traditional ML models such as Decision Trees (DT), Random Forests (RF), Support Vector Machines (SVM), and Naive Bayes (NB) were implemented, along with Tamil BERT and Malayalam BERT for the respective languages. These models were assessed using standard metrics.

From Table 2, we observe that the FFN classifier with BERT embeddings outperforms other models in both Tamil and Malayalam, achieving the highest accuracy of 95.68% and 88.75%, respectively. This demonstrates the effectiveness of transformer-based embeddings in capturing the complex linguistic structures of these languages. While traditional machine learning models with TF-IDF and BoW embeddings perform adequately, they lag behind deep learning approaches. Among traditional models, the RF classifier performs better for Tamil, while NB shows relatively stronger results for Malayalam.

However, both remain inferior to the FFN-BERT model, highlighting the advantage of deep contextualized embeddings in handling the linguistic complexities of Tamil and Malayalam language. The code for implementing this experiment can be found on GitHub.

6 Conclusion

This study focuses on detecting AI-generated product reviews in Tamil and Malayalam using transformer models, specifically Tamil-BERT and Malayalam-BERT, in addition to traditional ML approaches. The BERT models outperformed traditional ML models. Robust preprocessing techniques and accessible datasets form a solid foundation for identifying AI-generated content in low-resource languages. This framework enhances the credibility of user-generated reviews and supports NLP resource development, advancing research in the identification of AI-generated reviews across Tamil and Malayalam languages. The model achieves 95.68% accuracy on Tamil and 88.75% on Malayalam datasets.

Limitations

This study faces limitations due to the small dataset size for both languages, which may impact model performance. As low-resource languages, Tamil and Malayalam have limited representation of offensive and misogynistic words in available corpora, which constrains the effectiveness of BERT models. Additionally, models like XLM-RoBERTa and IndicBert, trained on significantly larger datasets, with more tokens and parameters than BERT-based models, could offer improved results, especially for mixed-code texts. To overcome these limitations, future work will focus on expanding datasets, incorporating multilingual models, and enhancing linguistic diversity to improve AI-generated review detection in Dravidian languages.

Table 2: Comparison of the proposed model with other models

Classifier	TE	Tamil					Malayalam				
		Acc	**P**	**R**	**F1**	**F1$^{\text{Macro}}$**	**Acc**	**P**	**R**	**F1**	**F1$^{\text{Macro}}$**
NB	TF-IDF	0.8086	0.8261	0.7500	0.7862	0.8065	0.8187	0.8000	0.8500	0.8242	0.8185
	BOW	0.7963	0.8209	0.7237	0.7692	0.7934	0.8186	0.8000	0.8500	0.8243	0.8186
DT	TF-IDF	0.8334	0.8356	0.8026	0.8188	0.8323	0.6937	0.6867	0.7125	0.6993	0.6936
	BOW	0.8641	0.8552	0.8552	0.8552	0.8636	0.6875	0.7027	0.6500	0.6753	0.6870
SVM	TF-IDF	0.8765	0.8590	0.8816	0.8701	0.8762	0.7750	0.7895	0.7500	0.7692	0.7749
	BOW	0.8580	0.8442	0.8553	0.8496	0.8575	0.7563	0.7971	0.6875	0.7383	0.7551
RF	TF-IDF	0.8951	0.9041	0.8684	0.8859	0.8943	0.7812	0.7922	0.7625	0.7770	0.7811
	BOW	0.8704	0.8235	0.9210	0.8695	0.8703	0.7937	0.79012	0.8000	0.7950	0.7937
FFN	**BERT**	**0.9568**	**0.9568**	**0.9568**	**0.9568**	**0.9566**	**0.8875**	**0.8897**	**0.8875**	**0.8873**	**0.8873**

Abbreviations: TE – Text Embedding, Acc – Accuracy, P – Precision, R – Recall.

Acknowledgment

We would like to thank DravidianLangTech-2025 at NAACL for providing the dataset used in the shared task. Their contributions have been crucial in advancing research on Dravidian languages.

References

M. Abdedaiem, B. Othman, and N. Charrad. 2023. Few-shot learning for fake news detection in low-resource languages. *ResearchGate*.

M. H. Al-Adhaileh and F. W. Alsaade. 2022. Bidirectional long-short term memory (bilstm) networks for fake review detection: A comparative study. *Springer*.

A. Bala and P. Krishnamurthy. 2023. Transfer learning for fake news detection in dravidian languages. *ResearchGate*.

S. Barman and M. Das. 2023. Multimodal approaches for sentiment analysis and abusive language detection in tamil and malayalam. *ResearchGate*.

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. 2021. 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*.

R. Eduri Raja and A. Bala. 2023. Fake news detection in dravidian languages using transfer learning models. *ResearchGate*.

Stanford CS224N et al. 2023. Multitask fine-tuning with smoothness-induced adversarial regularization for nlp tasks. *ResearchGate*.

S. Kumar, P. S. Venugopala, and K. R. Rao. 2024. Term frequency and review regeneration model for identifying ai-generated peer reviews. *Springer*.

Asrita Venkata Mandalam and Yashvardhan Sharma. 2021. Sentiment analysis of Dravidian code mixed data. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*.

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.

Eduri Raja, Badal Soni, and Samir Kumar Borgohain. 2023. Fake news detection in dravidian languages using transfer learning with adaptive finetuning. *Engineering Applications of Artificial Intelligence*, 126.

Pradeep Kumar Roy, Snehaan Bhawal, and Chinnaudayar Navaneethakrishnan Subalalitha. 2022. Hate speech and offensive language detection in dravidian languages using deep ensemble framework. *Computer Speech Language*, 75:101386.

Malliga Subramanian, B Premjith, Kogilavani Shanmugavadivel, Santhia Pandiyan, Balasubramanian Palani, and Bharathi Raja Chakravarthi. 2025. Overview of the Shared Task on Fake News Detection 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.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*.