The_Deathly_Hallows@DravidianLangTech 2025: AI Content Detection in Dravidian Languages

Kogilavani Shanmugavadivel[1] **, Malliga Subramanian**[2]**,**
Vasantharan K[1]**, Prethish G A**[1]**, Vijayakumaran S**[1]
[1]Department of AI, Kongu Engineering College, Perundurai, Erode.
[2]Department of CSE, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{vasantharank.ncc, prethish0409, vijayakumaran2k3}@gmail.com

Abstract

The DravidianLangTech@NAACL 2025 shared task focused on Detecting AI-generated Product Reviews in Dravidian Languages, aiming to address the challenge of distinguishing AI-generated content from human-written reviews in Tamil and Malayalam. As AI generated text becomes more prevalent, ensuring the authenticity of online product reviews is crucial for maintaining consumer trust and preventing misinformation. In this study, we explore various feature extraction techniques, including TF-IDF, Count Vectorizer, and transformer-based embeddings such as BERT-Base-Multilingual-Cased and XLM-RoBERTa-Large, to build a robust classification model. Our approach achieved F1-scores of 0.9298 for Tamil and 0.8797 for Malayalam, ranking 8th in Tamil and 11th in Malayalam among all participants. The results highlight the effectiveness of transformer-based embeddings in differentiating AI-generated and human-written content. This research contributes to the growing body of work on AI-generated content detection, particularly in underrepresented Dravidian languages, and provides insights into the challenges unique to these languages.

.

1 Introduction

With the rapid advancement of natural language generation models, AI-generated text has become increasingly prevalent across various domains, including online product reviews. While these models enhance automation and efficiency, they also raise concerns regarding the authenticity and reliability of online content. AI-generated reviews have the potential to influence consumer decisions, mislead potential buyers, and distort market perceptions. This issue is particularly critical in low-resource languages such as Tamil and Malayalam, where limited research has been conducted on detecting AI-generated content Premjith et al. 2025.

To address this challenge, the Shared Task on Detecting AI-generated Product Reviews in Dravidian Languages was organized, providing a dataset consisting of both human-written and AI-generated reviews in Tamil and Malayalam. Participants were tasked with developing models to accurately classify these reviews while considering the unique linguistic complexities of Dravidian languages. The evaluation metric used in this task was the F1-score, ensuring a balanced and robust assessment of model performance.

In this work, we present our methodology and findings in tackling this problem. We explored both traditional feature extraction methods, such as TF-IDF and Count Vectorizer, and transfer learning models, including BERT-Base-Multilingual-Cased and XLM-RoBERTa-Large. Through extensive experimentation and analysis, we highlight the challenges involved in distinguishing AI-generated reviews from human-authored ones and assess the effectiveness of various feature extraction techniques and classification models. Our results provide insights into the applicability of transformer-based embeddings for AI-generated content detection in underrepresented Dravidian languages.

2 Literature Review

The study conducted by Wu et al. 2023 examined the role of ChatGPT in credit default prediction by comparing AI-generated and human-generated loan assessments. Their findings indicated that ChatGPT-generated insights contributed to improved predictive accuracy, underscoring the model's potential in financial decision-making. Agrawal et al. 2019 investigated the impact of AI on human labor, particularly differentiating between prediction and judgment tasks. The study demonstrated that AI significantly reduced predic-

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 292–296
May 3, 2025 ©2025 Association for Computational Linguistics

tion costs, leading to increased variance in outcomes and altering the perceived value of human judgment in decision-making processes. Molina and Sundar 2024 explored the factors influencing user trust in AI-driven content moderation. Their research found that individuals with lower trust in human moderators were more inclined to favor AI-based moderation systems, providing insights into the evolving trust dynamics between automation and human decision-making.

Cao et al. 2023 conducted a comprehensive survey on AI-generated content (AIGC), tracing its evolution from early generative adversarial networks (GANs) to advanced models such as ChatGPT and DALL-E-2. The study outlined significant advancements, identified key challenges, and discussed future directions for generative AI applications. Singhal and Bedi 2024 presented a transformer-based approach for Tamil code-mixed sentiment analysis, specifically applied to hate speech detection. Their ensemble model achieved the highest ranking at LT-EDI 2024, demonstrating the effectiveness of RoBERTa-based architectures in multilingual and code-mixed settings. Devananthan and Nair 2023 examined multilingual sentiment analysis on Indian Twitter, evaluating a range of machine learning and deep learning models. Their research contributed to the development of a robust framework capable of processing diverse linguistic content efficiently. Kumaresan et al. 2022 focused on hope speech detection in code-mixed Tamil, English, and Malayalam. The study employed transformer-based models that achieved competitive F1 scores, highlighting their effectiveness in sentiment classification and content moderation tasks. Li et al. 2021 proposed a cross-lingual named entity recognition (NER) approach utilizing XLM-RoBERTa in conjunction with parallel corpora. Their approach enhanced entity alignment without the need for direct translation, surpassing the performance of unsupervised methods across multiple languages. Gaikwad et al. 2023 conducted a comparative analysis of multilingual sentiment analysis models. Their findings revealed that XLM-RoBERTa attained the highest accuracy (78.37%), demonstrating its adaptability and efficacy across various linguistic contexts. Raja et al. 2023 focused on fake news detection in Malayalam by optimizing an XLM-RoBERTa model. Their model achieved a macro-averaged F-score of 87%, securing the second rank in the DravidianLangTech competition, further reinforcing the effectiveness of transformer-based models in tackling misinformation detection tasks.

3 Dataset Description

The dataset for this task consists of AI-generated and human-written product reviews in two Dravidian languages: Malayalam and Tamil. The dataset was created to support the development of models capable of distinguishing between machine-generated and human-authored content in online reviews .

The Tamil dataset comprises 808 samples, with 405 AI-generated reviews and 403 human-written reviews. The Malayalam dataset contains 800 samples, evenly split between AI-generated and human-written reviews. Each sample is a textual review, and the datasets provide a balanced distribution to ensure fair model evaluation.

Dataset	AI-Generated	Human
Tamil	405	403
Malayalam	400	400

Table 1: Distribution of AI-generated and human-written reviews in the dataset.

4 Task Description

To tackle the increasing difficulty of identifying AI-generated material in online reviews, the Shared Task on Detecting AI-Generated Product Reviews in Dravidian Languages (DravidianLangTech@NAACL 2025) was created. As AI models get more complex, it is essential to differentiate between evaluations written by humans and those written by robots in order to preserve credibility and confidence in online markets. In Malayalam and Tamil, participants will create and assess models that categorise product reviews as either human-written or AI-generated. For training and testing, the dataset will be made available in an organised manner. Participants can download the dataset and submit their models for review on CodaLab.

The F1-score, a commonly used statistic for classification tasks in NLP, will be used to evaluate the model's performance. Participants from a variety of academic disciplines are invited to participate in this shared challenge, which aims to improve the recognition of AI-generated material in low-resource languages.

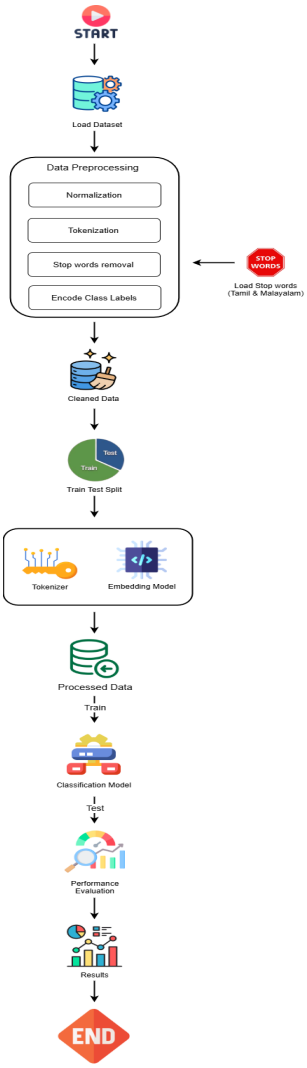

Figure 1: Proposed Model Workflow

5 Methodology

5.1 Preprocessing Dataset

The collection includes both human-written and AI-generated product reviews in Malayalam and Tamil. An ID, a textual review (transcript), and a class label specifying whether the review was prepared by a person or by artificial intelligence make up each instance.

Because Malayalam stopwords were not available in Advertools, stopwords for Malayalam were retrieved from a publicly accessible Git repository, whereas stopwords for Tamil were eliminated using the Advertools library. Tokenisation, stopword elimination, and text normalisation were among the preparation procedures. The *Indic NLP Library* was used to normalise the Tamil text, and the *indic-tokenize* package was used to tokenise it. After preprocessing, the text was saved for further use.

5.2 Feature Extraction

To obtain meaningful text representations, three different feature extraction techniques were used:

1. **TF-IDF:** Term Frequency-Inverse Document Frequency (TF-IDF) was used to convert text into numerical vectors.

2. **Count Vectorizer:** This method created a sparse representation of word occurrences in the corpus.

3. **Transformer-Based Embeddings:** Two pre-trained transformer models, BERT-Base-Multilingual-Cased and XLM-RoBERTa-Large, were used to extract contextual embeddings. The models were loaded using the Hugging Face transformers library.

For transformer-based embeddings, tokenization was performed using the corresponding model's tokenizer. The processed text was converted into tokenized sequences with a maximum length of 512 tokens. The embeddings were extracted by taking the mean of the last hidden state outputs of the model.

5.3 Classification Model

A deep learning-based classification model was developed to distinguish AI-generated reviews from human-written reviews. The model architecture included:

- A fully connected dense layer with 256 neurons and ReLU activation.
- Batch normalization and dropout (0.5) to prevent overfitting.
- Another dense layer with 128 neurons and ReLU activation, followed by batch normalization and dropout (0.5).
- A final dense output layer with softmax activation for classification.

The model was trained using the Adam optimizer with categorical cross-entropy loss. The dataset was split into 80% training and 20% testing using the *train_test_split* function from Scikit-learn.

6 Limitations

This study faces two primary limitations. First, the high computational cost of transformer-based models like XLM-RoBERTa-Large and BERT-Base-Multilingual-Cased makes them resource-intensive,

Models Used	Tamil	Malayalam
BERT-Base-Multilingual-Cased	94%	94%
TF-IDF	81%	76%
Count Vectorizer	83%	76%
XLM-RoBERTa-Large	96%	94%

Table 2: Performance of Different Models in Tamil and Malayalam in text

limiting their accessibility to researchers with constrained computational resources. Second, the dataset size is relatively small, with only 808 reviews in tamil and 800 in malayalam, which may affect the model's ability to generalize across different domains such as social media or news articles. Expanding the dataset and optimizing models for efficiency would enhance the applicability of AI content detection in Dravidian languages.

7 Performance Evaluation

Using the Accuracy, we assessed several text representation methods and classification models. Conventional techniques such as Count Vectorizer and TF-IDF shown difficulties in capturing contextual semantics, achieving accuracy of 76% for Malayalam and 81% and 83% for Tamil.

Transformer-based models performed noticeably better than conventional methods. While XLM-RoBERTa-Large performed the best, achieving 96% for Tamil and 94% for Malayalam, BERT-Base-Multilingual-Cased had an accuracy of 94% for both languages. These findings demonstrate how well deep contextual embeddings work to differentiate between writing produced by AI and text authored by humans.

8 Conclusion

This study explored various feature extraction techniques and classification models to distinguish AI-generated and human-written product reviews in Tamil and Malayalam. Traditional methods such as TF-IDF and Count Vectorizer were found to be less effective due to their inability to capture contextual semantics. Transformer-based models, particularly XLM-RoBERTa-Large, provided the highest accuracy, demonstrating the effectiveness of deep contextual embeddings. The results emphasize the importance of using pre-trained multilingual models for low-resource languages. By leveraging transformer-based architectures, we achieved an Accuracy of 96 for Tamil and 94 for Malayalam, outperforming traditional statistical approaches. Future research can explore fine-tuning transformer models on larger domain-specific datasets and incorporating additional linguistic features to further enhance classification accuracy. Additionally, integrating explainable AI techniques could provide insights into model decision-making, making AI-generated content detection more interpretable and trustworthy. The source code for our approach is available at https://github.com/vasantharan/Detecting_AI_generated_product_reviews_in_Dravidian_languages.

References

Ajay Agrawal, Joshua S Gans, and Avi Goldfarb. 2019. Exploring the impact of artificial intelligence: Prediction versus judgment. *Information Economics and Policy*, 47:1–6.

Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. 2023. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226*.

AG Devananthan and Lekshmi S Nair. 2023. Exploring multilingual indian twitter sentiment analysis: A comparative study. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–8. IEEE.

Arya Gaikwad, Pranav Belhekar, and Vinayak Kottawar. 2023. Advancing multilingual sentiment understanding with xgboost, svm, and xlm-roberta. In *International Conference on Data Science, Machine Learning and Applications*, pages 990–1000. Springer.

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Elizabeth Sherly, Sangeetha Sivanesan, and Bharathi Raja Chakravarthi. 2022. Transformer based hope speech comment classification in code-mixed text. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 120–137. Springer.

Bing Li, Yujie He, and Wenjin Xu. 2021. Cross-lingual named entity recognition using parallel corpus: A new approach using xlm-roberta alignment. *arXiv preprint arXiv:2101.11112*.

Maria D Molina and S Shyam Sundar. 2024. Does distrust in humans predict greater trust in ai? role

of individual differences in user responses to content moderation. *New Media & Society*, 26(6):3638–3656.

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 Sami Kumar Borgohain. 2023. nlpt malayalm@ dravidianlangtech: Fake news detection in malayalam using optimized xlm-roberta model. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 186–191.

Kriti Singhal and Jatin Bedi. 2024. Transformers@ dravidianlangtech-eacl2024: Sentiment analysis of code-mixed tamil using roberta. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 151–155.

Zongxiao Wu, Yizhe Dong, Yaoyiran Li, and Baofeng Shi. 2023. Unleashing the power of text for credit default prediction: Comparing human-generated and ai-generated texts. *Available at SSRN 4601317*.