CUET_NLP_FiniteInfinity@DravidianLangTech 2025: Exploring Large Language Models for AI-Generated Product Review Classification in Malayalam

Md. Zahid Hasan, Safiul Alam Sarker, MD Musa Kalimullah Ratul
Kawsar Ahmed and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
{u1904099, u1904041, u1904071, u1804017}@student.cuet.ac.bd
moshiul_240@cuet.ac.bd

Abstract

Detecting AI-generated product reviews is a critical challenge in natural language processing (NLP), particularly for low-resource languages like Malayalam. In this study, we propose a large language model (LLM)-based approach to identify AI-generated reviews in Dravidian languages, focusing on the product review domain. We systematically evaluated multiple LLMs on a dedicated Malayalam dataset to assess their effectiveness in distinguishing between human-written and AI-generated reviews. Our experiments demonstrate that the Gemma-2B model outperforms other models, achieving a macro F1-score of 89.99%.Our approach secured 5^{th} place in the DravidianLangTech@NAACL 2025 shared task for Malayalam, highlighting the potential of LLMs in tackling the challenges of AI-generated review detection in low-resource languages. Our findings highlight the potential of LLMs in detecting AI-generated content in underrepresented languages, contributing to advancements in Dravidian language processing and the broader field of AI-generated content identification.

1 Introduction

The detection of AI-generated product reviews is an emerging challenge in natural language processing (NLP), particularly in the context of low-resource languages. As AI-generated text becomes increasingly sophisticated, distinguishing between human-written and machine-generated content is crucial for ensuring the authenticity of online reviews. In e-commerce platforms, product reviews play a pivotal role in shaping consumer trust, influencing purchasing decisions, and maintaining credibility between buyers and sellers. The proliferation of AI-generated reviews, however, poses a significant risk to the reliability of these platforms, making automated detection systems essential.

While extensive research has been conducted on AI-generated text detection in English (Salminen et al., 2022; Luo et al., 2023), there remains a notable gap in studies focusing on particularly Malayalam language. The limited availability of high-quality annotated datasets, coupled with the linguistic complexity of these languages, presents significant challenges in building robust detection models. This shared task addresses these gaps by encouraging the development of effective machine learning approaches for detecting AI-generated product reviews in Malayalam arranged by DravidianLangTech@NAACL 2025 (Premjith et al., 2025).

In this study, AI-generated product review detection is formulated as a binary classification problem, where the goal is to classify a given review as either human-written or AI-generated. To tackle the challenges associated with low-resource languages, we explore the application of large language models (LLMs) fine-tuned specifically for this task. Our contributions are as follows:

- We propose a fine-tuned large language model designed for AI-generated product review detection in Malayalam language.
- We systematically evaluate multiple LLMs including Gemma-2-2b (Team et al., 2024), Llama-3.2-3B (AI, 2024), sarvam-1[1], Qwen2.5-3B (Yang et al., 2024), and BharatGPT-3B-Indic[2] to identify the most effective approach for this task.

This work aims to advance the field of AI-generated text detection in low-resource language and establish a strong foundation for future research in this domain.

[1] https://huggingface.co/sarvamai/sarvam-1
[2] https://huggingface.co/CoRover/BharatGPT-3B-Indic

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 651–656

2 Related Work

The growth of AI-generated content, especially in the form of product evaluations, has emerged as a serious concern in e-commerce and social media platforms. Recent research have emphasised the increased competence of AI models in creating human-like writing, making it more difficult to discern between genuine and false evaluations (Luo et al., 2023). Gambetti and Han (2023) suggested utilizing AI to combat machine-generated phony reviews, getting an F1 score of 0.92 on a restaurant review data set using ensemble learning and contextual embeddings. Similarly, Birim et al. (2022) applied topic modeling approaches to detect patterns suggestive of bogus reviews, reporting an accuracy of 89.5% on a multilingual dataset. Salminen et al. (2022) examined the development and detection of fake reviews, attaining an F1 score of 0.87 by integrating language characteristics and behavioral analysis. Shibani et al. (2024) examined generative AI for Tamil writing help, getting an F1 score of 0.89 in recognising AI-generated text. De et al. (2021) suggested a transformer-based technique for multilingual false news identification, obtaining 87.3% accuracy and an F1 score of 0.85 using mBERT. Budhi et al. (2021) handled unbalanced datasets in fake review identification, reaching 91.2% accuracy and an F1 score of 0.88 by resampling and textual characteristics. Cheng et al. (2024) used graph neural networks (GNNs) to detect bogus reviewers, attaining an F1 score of 0.92 by assessing social context. Mukherjee (2024) emphasises on avoiding AI-generated fraud in marketing, underlining the importance for robust detection techniques. These findings underscore the significance of language-specific and context-aware models, particularly for low-resource languages like Malayalam and indicate the promise of advanced techniques like transformers and GNNs in fighting AI-generated fraudulent information. These findings together underline the necessity for powerful, language-specific detection approaches, especially for low-resource languages like Dravidian languages, to face the rising threat of AI-generated bogus reviews.

3 Dataset and Task Description

The shared task on "Detecting AI-generated Product Reviews in Dravidian Languages" (Premjith et al., 2025) focuses on identifying AI-generated and human-written reviews in Malayalam. With the increasing sophistication of AI tools, this task addresses the need for accurate detection models in the domain of online reviews, where authenticity is critical.

Participants were provided datasets comprising human-written and AI-generated reviews. As shown in Table 1, the training set includes 800 reviews, while the test set contains 200 reviews. The task invites global participation via CodaLab[3] to enhance AI detection for Dravidian language.

Set	Class	S_C	W_T	W_U	Avg. Len
Train	HUMAN	400	6174	3357	15.44
	AI	400	4121	2317	10.30
Test	HUMAN	100	2027	1462	20.27
	AI	100	1053	821	10.53

Table 1: Dataset Statistics: Sentence Count (S_C), Total Words (W_T), Unique Words (W_U), and Average Length (Avg. Len)

4 System Overview

In this study, we investigate a comprehensive suite of approaches—including machine learning (ML), deep learning (DL), transformer-based methods, and large language models (LLMs)—to detect AI-generated product reviews in Malayalam. Figure 1 presents a schematic overview of our proposed methodology. Detailed implementation and source code for the system are available on GitHub[4].

4.1 Machine Learning Approaches

We evaluated traditional ML models, including Logistic Regression, Support Vector Machines (SVM), Random Forest (RF), Naïve Bayes (NB), Decision Trees (DT), Kernel SVM, and Stochastic Gradient Descent (SGD), for product review classification. Textual data was transformed into high-dimensional vectors using TF-IDF (Takenobu, 1994) and CountVectorizer, with TF-IDF limited to 1000 features for optimal performance. Logistic regression used a maximum of 1000 iterations. Both linear and kernel SVMs were tested, with the RBF kernel applied for non-linear classification and regularization parameter $C = 0.80$ for linear SVM.

4.2 Deep Learning Approaches

We explored several deep learning models, including CNN, BiLSTM, BiLSTM+Attention,

[3] https://codalab.org/

[4] https://github.com/zahid99hasan/AI-Generated-Text-Detection

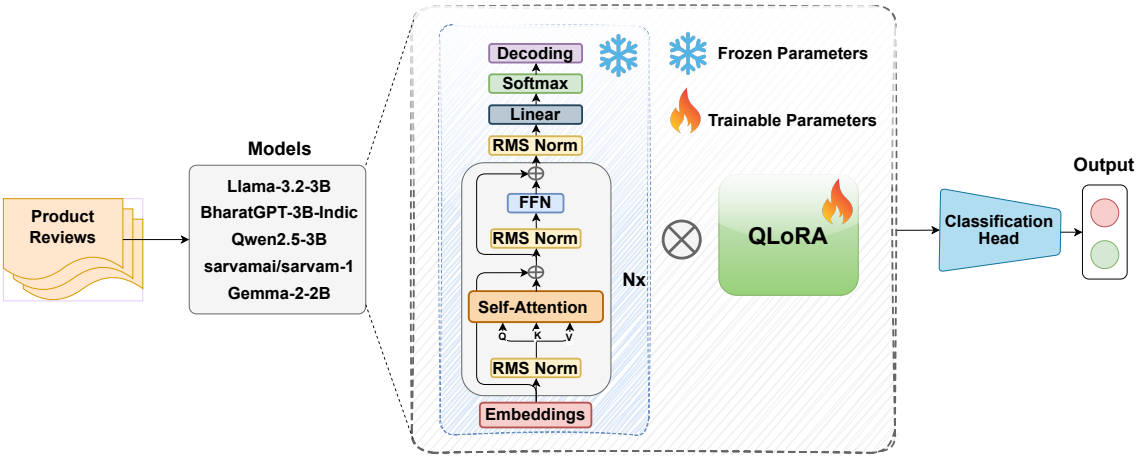

Figure 1: Schematic process for product review detection.

Keras+CNN, GloVe+CNN, and GloVe+BiLSTM. The CNN model utilized 128 filters with a kernel size of 5 and 24 neurons in the dense layer. In the BiLSTM+Attention model, ReLU was used in the dense layer and Sigmoid in the output layer. BiLSTM (Schuster and Paliwal, 1997) enhanced performance by capturing bidirectional contextual information. Keras+CNN and GloVe+CNN employed 1D convolutional layers for processing sequential data. GloVe+BiLSTM used an LSTM architecture with 64 neurons to effectively capture long-term dependencies in text sequences.

4.3 Transformer-Based Approaches

Several pre-trained transformer models from Hugging Face were leveraged for product review classification, including mBERT (Devlin, 2018), XLM-R (Conneau, 2019), MalayalamBERT (Joshi, 2022), and IndicBERT (Kakwani et al., 2020). Before passing data through the transformers, preprocessing and tokenization were performed. All models were trained using a learning rate of 5×10^{-5}, a batch size of 16 for both training and validation, and 4 epochs to achieve optimal results.

4.4 LLM-Based Approaches

Large language models (LLMs) with efficient fine-tuning via QLoRA (Quantized Low-Rank Adaptation) (Dettmers et al., 2024) were utilized for classifying product reviews in Dravidian languages (e.g., Malayalam) as Human-generated or AI-generated. Pre-trained LLMs, including Llama-3.2-3B (AI, 2024), BharatGPT-3B-Indic[5], Qwen2.5-3B (Yang et al., 2024), sarvamai/sarvam-1[6], and Gemma-2-2B (Team et al., 2024), were employed. QLoRA preserves base model parameters while introducing trainable low-rank adapters, enabling efficient task-specific adaptation. Figure 1 illustrates the proposed approach. The classification pipeline begins with tokenizing input reviews, which are then processed through frozen LLM layers to generate contextual embeddings. These embeddings are modified by QLoRA adapters, which introduce trainable low-rank updates while keeping the base model parameters unchanged. The adapted embeddings are then passed to a classification head, which performs the final prediction to determine whether the review is Human-generated or AI-generated. This approach efficiently adapts pre-trained LLMs for the classification task while maintaining computational efficiency and robust performance. The proposed model (Gemma-2-2B) was trained for 10 epochs with a batch size 16 and a learning rate of 1e-4, achieving the best overall performance. QLoRA was applied with a rank of 4, alpha of 16, a dropout rate of 0.1, and no bias.

5 Results and Analysis

Table 2 demonstrates the evaluation results large language models on the test set.

Results revealed that Gemma-2-2B for Malayalam earned the most elevated macro F1-score (89.99%) among the LLM approaches. On the other hand, for Malayalam sarvam-1 with macro F1-score (84.47%) surpasses all the models except Gemma-2-2B among the models. For Malay-

[5] https://huggingface.co/CoRover/BharatGPT-3B-Indic

[6] https://huggingface.co/sarvamai/sarvam-1

alam, Llama-3.2-3B perform poorly with the lowest macro F1 score.

ML Models			
Classifier	**G-mean(%)**	**F1(%)**	**Ac(%)**
LoR	64.00	66.00	67.00
SVM	63.00	66.00	67.00
RF	61.00	65.00	65.00
NB	59.00	62.00	62.00
DT	59.00	57.00	57.00
KerneL SVM	63.00	66.00	67.00
SGD	**67.00**	**69.00**	**69.00**
DL Models			
Classifier	**G-mean(%)**	**F1(%)**	**Ac(%)**
CNN	72.74	73.20	73.50
BiLSTM	64.99	66.66	68.00
BiLSTM + Attention	17.32	36.58	51.50
Keras + CNN	73.99	73.99	74.00
GloVe + CNN	66.11	68.32	70.00
GloVe + BiLSTM	**74.00**	**74.00**	**74.00**
Transformers			
Classifier	**G-mean(%)**	**F1(%)**	**Ac(%)**
mBERT	**74.90**	**78.62**	**78.75**
IndicBERT	64.95	74.08	75.00
XLM-R	54.56	63.16	64.37
MalayalamBERT	70.57	77.66	78.12
LLMs			
Classifier	**G-mean(%)**	**F1(%)**	**Ac(%)**
Gemma-2-2B	**90.06**	**89.99**	**90.00**
sarvam-1	84.63	84.47	84.50
Llama-3.2-3B	35.35	33.33	50.00
Qwen2.5-3B	25.97	37.60	40.50
BharatGPT-3B-Indic	30.87	45.50	29.60

Table 2: Performance of the different methods on the test set

LLMs perform well on the validation set, which is splinted from train set. Over fitting can be a reason for that. BhartGPT-3B-Indic expected to perform really well, but in reality it shows an average performance. This encourage to explore more models for better performance. To explore better performance, this work explored Qwen2.5-3B, Llama-3.2-3B models as well.

5.1 Error Analysis

A comprehensive error analysis is performed to offer in-depth insights into the performance of the proposed model.

Quantitative Analysis

Since the gold labels for the test set were disclosed, Figure 2 presents the confusion matrix that categorizes product reviews into Human and AI predicted. The figure indicates that out of 200 reviews , 180 were accurately predicted. Al reviews (13) predicted more incorrectly compare to Human reviews (7), this occurred because AI can generate humanoid reviews. And for short-length data (less than or equal to 10 words), our proposed model perform better than large length of data. The proposed models only trained on the given dataset.

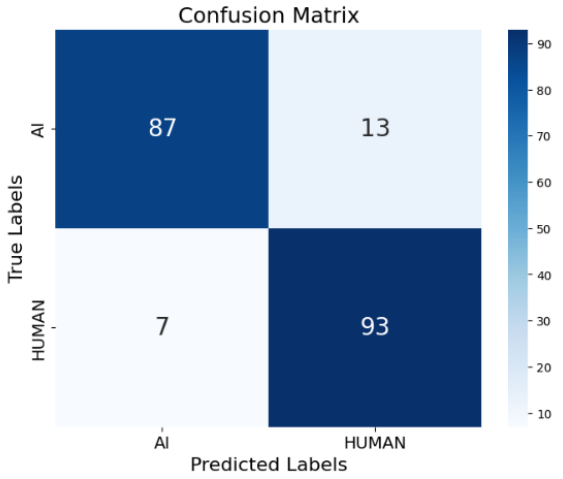

Figure 2: Confusion matrix of Gemma-2-2B model.

Qualitative Analysis

Table 3 presents some predicted outputs of the tested model. In the first, second and fourth data, the model successfully predicted the review of the data. On the other hand, it failed to do so in the third data. The Gemma-2-2B model for Malayalam, which is fine-tuned in this work, is primarily trained on given datasets from different sources; this could be one of the reasons for these model's failure in some Malayalam data. Furthermore, incorrect predictions arose due to stylistic similarities between formal human-authored text and AI-generated patterns, insufficient diversity in the training dataset, topical overlap with AI-prevalent themes, limitations in model capacity from low-rank adaptation and quantization, and loss of context from input truncation.

Text Sample	Actual	Predicted
Sample1: കോവയ്ക്ക ഉപ്പിലിട്ടത് എൻറെ ജീവിതത്തിൽ ഇതുവരെ കഴിച്ചിട്ടില്ല (I have never eaten salted cod in my life.)	Human	Human
Sample2: കോവയ്ക്ക ഉപ്പിലിട്ടത് ഞാൻ ഇതുവരെ കഴിച്ചിട്ടില്ല, കഴിക്കാൻ മനസില്ല. (I have never eaten salted cod, and I don't feel like eating it.)	AI	AI
Sample3: ഇക്കാലത്ത് ഒക്കെ ടാറ്റയുടെ ഡിസൈൻ കണ്ട് കണ്ണിന്ന ഊട്ടുപോലും ഇട്ടുന്നണ്ട് (Nowadays, Tata's designs are even making my eyes water.)	AI	Human
Sample4: കാൽ കാശിന്ന കൊള്ളാത്ത ഭക്ഷണമാണ് ചേട്ടാ (The food is not worth the money, brother.)	Human	Human

Figure 3: Sample predictions with actual and predicted reviews

6 Conclusion

This study explored the effectiveness of several large language models in detecting AI-generated content within a Malayalam product review dataset. Our findings indicate that the Gemma-2-2B model excelled in this task, achieving a macro F1-score of 89.99%. The results underscore the potential of transformer-based approaches for this application and motivate further exploration of alternative transformer architectures and LLMs to enhance performance.

Limitations

Despite the promising results, our study has several limitations. First, the relatively small dataset may not fully capture the diversity of both AI-generated and human-written reviews, potentially limiting the generalizability of our findings. Second, our study focuses exclusively on Malayalam, restricting the applicability of the approach to other Dravidian and low-resource languages. Additionally, while some large language models, such as Gemma-2B, performed well, others underperformed, highlighting the need for further investigation into model selection and optimization for AI-generated text detection. A key challenge observed is the multilingual incapability of certain LLMs, which may stem from insufficient training data in Dravidian languages. Finally, the dataset may not encompass the full spectrum of AI-generated writing styles, which could affect the robustness of the classification models in real-world scenarios. To address these limitations, future work should explore more extensive and diverse datasets, use models with more parameters, extend the approach to other Dravidian languages, and refine model architectures to enhance performance in low-resource multilingual settings.

Acknowledgments

We thank the DravidianLangTech 2025 shared task organizers for running this task. This work was supported by the Directorate of Research & Extension (DRE), Chittagong University of Engineering & Technology (CUET).

References

Meta AI. 2024. Llama 3.2: Revolutionizing edge ai and vision with open-source models. Accessed: 2025-01-29.

Şule Öztürk Birim, Ipek Kazancoglu, Sachin Kumar Mangla, Aysun Kahraman, Satish Kumar, and Yigit Kazancoglu. 2022. Detecting fake reviews through topic modelling. *Journal of Business Research*, 149:884–900.

Gregorius Satia Budhi, Raymond Chiong, and Zuli Wang. 2021. Resampling imbalanced data to detect fake reviews using machine learning classifiers and textual-based features. *Multimedia Tools and Applications*, 80:13079–13097.

Li-Chen Cheng, Yan Tsang Wu, Cheng-Ting Chao, and Jenq-Haur Wang. 2024. Detecting fake reviewers from the social context with a graph neural network method. *Decision Support Systems*, 179:114150.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Arkadipta De, Dibyanayan Bandyopadhyay, Baban Gain, and Asif Ekbal. 2021. A transformer-based approach to multilingual fake news detection in low-resource languages. *Transactions on Asian and Low-Resource Language Information Processing*, 21(1):1–20.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alessandro Gambetti and Qiwei Han. 2023. Combat ai with ai: Counteract machine-generated fake restaurant reviews on social media. *arXiv preprint arXiv:2302.07731*.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Jiwei Luo, Guofang Nan, Dahui Li, and Yong Tan. 2023. Ai-generated review detection. *Available at SSRN 4610727*.

Anirban Mukherjee. 2024. Safeguarding marketing research: The generation, identification, and mitigation of ai-fabricated disinformation. *arXiv preprint arXiv:2403.14706*.

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.

Joni Salminen, Chandrashekhar Kandpal, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J Jansen. 2022. Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services*, 64:102771.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Antonette Shibani, Faerie Mattins, Srivarshan Selvaraj, Ratnavel Rajalakshmi, and Gnana Bharathy. 2024. Tamil co-writer: Towards inclusive use of generative ai for writing support. In *LAK Workshops*, pages 240–248.

Tokunaga Takenobu. 1994. Text categorization based on weighted inverse document frequency. *Information Processing Society of Japan, SIGNL*, 94(100):33–40.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.