CUET_Absolute_Zero@DravidianLangTech 2025: Detecting AI-Generated Product Reviews in Malayalam and Tamil Language Using Transformer Models

Anindo Barua, Sidratul Muntaha, Momtazul Arefin Labib, Samia Rahman, Udoy Das[†], Hasan Murad
Department of Computer Science and Engineering,
Chittagong University of Engineering and Technology, Bangladesh
[†]East Delta University, Bangladesh
{u2004040, u2004041, u1904111, u1904022}@student.cuet.ac.bd,
udoy.d@eastdelta.edu.bd, hasanmurad@cuet.ac.bd

Abstract

Artificial Intelligence (AI) is opening new doors of learning and interaction. However, it has its share of problems. One major issue is the ability of AI to generate text that resembles human-written text. So, how can we tell apart human-written text from AI-generated text? With this in mind, we have worked on detecting AI-generated product reviews in Dravidian languages, mainly Malayalam and Tamil. The "Shared Task on Detecting AI-Generated Product Reviews in Dravidian Languages," held as part of the DravidianLangTech Workshop at NAACL 2025 has provided a dataset categorized into two categories, human-written review and AI-generated review. We have implemented four machine learning models (Random Forest, Support Vector Machine, Decision Tree, and XGBoost), four deep learning models (Long Short-Term Memory, Bidirectional Long Short-Term Memory, Gated Recurrent Unit, and Recurrent Neural Network), and three fine-tuned transformer-based on their performance in detecting AI-generated text (AI-Human-Detector, Detect-AI-Text, and E5-Small-Lora-AI-Generated-Detector). We have conducted a comparative study among all the models by training and evaluating each model on the dataset. We have discovered that the transformer, E5-Small-Lora-AI-Generated-Detector, has provided the best result with an F1 score of 0.8994 on the test set ranking 7th in the Malayalam language. Tamil has a higher token overlap and richer morphology than Malayalam. Thus, we obtained a worse F1 score of 0.5877 ranking 28th position in the Tamil language among all participants in the shared task.

1 Introduction

AI-generated content has created a significant challenge in distinguishing authenticity. The misuse of AI can even lead to the spread of misinformation. For example, online product reviews that influence consumer decision-making are now raising concerns about their trustworthiness. A significant amount of previous studies have been conducted on AI-generated product review detection. The majority of works have been done in high-resource languages, like English (Mikros et al., 2023,Abburi et al., 2023,Valdez-Valenzuela et al., 2024,Marchitan et al., 2024). But, little has been done for low-resource languages Malayalam and Tamil. Difficult lexemes and no specific pattern for tokens make the detection of AI-generated Malayalam and Tamil product reviews difficult. Traditional Machine-Learning models have been found to struggle with the contextual awareness and linguistic complexities of Tamil and Malayalam (Islam et al., 2023). Likewise, Deep Learning-based RNN models such as GRU, LSTM, and BiLSTM have been found to struggle with capturing long-range dependencies and contextual nuances. (Gaggar et al., 2023)

Despite the demonstrated accuracy and widespread adoption in various natural language processing tasks, the Transformer-based approaches have not been utilized for AI-generated product review detection in Tamil and Malayalam language. This brings us to the primary objective of our paper, which is detecting AI-generated product reviews in Dravidian languages, mainly in Malayalam and Tamil, using transformer models. The "Shared Task on Detecting AI-generated Product Reviews in Dravidian Languages", at NAACL 2025 has provided a balanced yet limited dataset categorized into two categories, human-written reviews and AI-generated reviews.

We have implemented four machine learning models (Random Forest, SVM, Decision Tree, and XGboost), four deep learning models (RNN, GRU, LSTM, and BilSTM), and three transformer-based models (Ai-Human-Detector, Detect-Ai-Text, and E5-Small-Lora-Ai-Generated-Detector). We have conducted

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 476–481

a comparative study among the models by evaluating each model on the dataset.

The transformer-based (Alshammari et al., 2024, Mo et al., 2024) approaches that rely on self-attention mechanisms can capture long-range dependencies along with contextual connections within texts. Thus, languages like Malayalam and Tamil, having complex morphological and syntactic structures, are generalized well by transformer-based models rather than machine learning and deep learning models. Among the three fine-tuned transformer-based models, we have found that the transformer "E5-Small-Lora-Ai-Generated-Detector" has provided the best result. In the case of the Malayalam language, it has obtained an F1 score of 0.8994 on the test set, ranking 7th. For the Tamil language, due to a higher token overlap and richer morphology, it has ranked 28th with an F1 score of 0.5877 among all participants in the shared task (Premjith et al., 2025). The core contributions of this research work are:

- To implement and compare the traditional ML models, deep learning models, and transformer-based models.
- To handle insufficient data by augmentation, to conduct a detailed error analysis, and to investigate the causes of misclassification.

The implementation details have been provided in the following: GitHub Repository.

2 Related Work

The previous studies on AI-generated Text Detection can be categorized under Machine Learning, Deep Learning, and Transformer-Based approaches.

Traditional Machine Learning (ML) techniques have been applied for AI-generated Text Detection in online social media platforms (Gaggar et al., 2023). Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF) have been used widely for AI-generated text detection with SVM providing the best result (Cingillioglu, 2023).

In comparison, Deep learning-based approaches are less dependent on explicitly defined features as they learn patterns and features automatically. The models integrate various layers, including LSTM, Transformer, and CNN, to perform tasks such as text classification and sequence labeling. This combination allows the model to effectively capture linguistic patterns improving text detection capabilities (Mo et al., 2024).

Generative language models like BERT, RoBERTa, and GPT have been used in detecting AI-based techniques (Mikros et al., 2023). DistilBert-Base-Uncased Model, Detect-Ai-Text, And E5-Small-Lora-Ai-Generated-Detector have demonstrated their effectiveness in the field of AI-generated text detection. In addition, several shared tasks (Nguyen et al., 2023,Maloyan et al., 2022) are accelerating the research effort leading to greater refinement of the AI-generated text detection methods.

3 Dataset

In the shared task "Detecting AI-Generated Product Reviews in Dravidian Languages" at the DravidianLangTech Workshop at NAACL 2025, the dataset provided for AI-generated product reviews contains Malayalam and Tamil language reviews. Table 1 and Table 2 contain the Tamil and Malayalam train datasets respectively. Initially, the datasets are limited in size. We have used data augmentation via back translation using the substitution method to compensate for data scarcity. Post-augmentation, the Tamil training split expanded to 806 human-written and 810 AI-generated reviews, while the Malayalam split expanded to 800 of each, providing more data for improved model training.

Table 1: Category-wise distribution in the Tamil dataset

Sets	AI	HUMAN	Total
Train	405	405	810
Development	405	401	806
Test	50	50	100
Total	860	856	1716

Table 2: Category-wise distribution in the Malayalam dataset

Sets	AI	HUMAN	Total
Train	400	400	800
Development	400	400	800
Test	100	100	200
Total	900	900	1800

4 Methodology

We have worked on a binary classification task involving low-resource languages to classify product reviews as AI-generated or human-written. At the outset, feature extraction has been performed. Multiple ML and DL algorithms were applied for analysis.

ML-based approaches, including Decision Tree, Support Vector Machine(SVM), Random Forest, and XGBoost have been used. SVM has been incorporated with a soft margin in the hyperplane. Deep Learning-based approaches, including RNN, LSTM, GRU, and BiLSTM have been used. This dataset is tokenized using NLTK's word-level tokenizer, which outputs a list of individual words and punctuation marks. We have applied Word2Vec for feature extraction due to its simplicity and ease of implementation. We could have replaced it with FastText, which handles out-of-vocabulary words better by leveraging subword information. Yet, we focused on transformer-based models for their superior performance in capturing complex linguistic patterns and the time constraints of the shared task. This decision not to incorporate FastText embeddings represents a limitation of our study.

Additionally, the system has been enhanced using different transformer models, as illustrated in Figure 1.

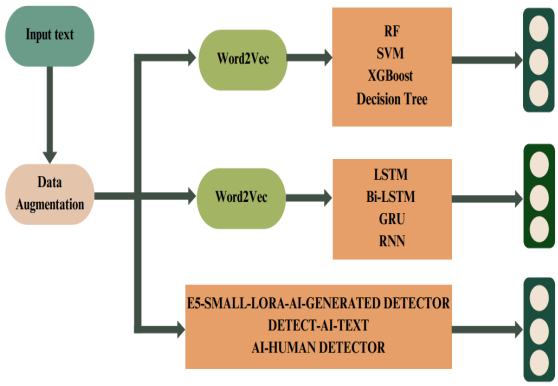


Figure 1: Abstract process of AI-generated product review detection

Three fine-tuned transformer-based models, E5-Small-LoRA-AI-Generated-Detector, AI-Human-Detector, and Detect-AI-Text, have been fine-tuned on back-translated Tamil and Malayalam text using the Trainer API of HuggingFace. Due to the limited data available in the dataset, we have implemented Synonym-based Augmentation using Contextual embedding (Pavlyshenko and Stasiuk, 2023). At first, we have used the Google-trans library where the text data has been back-translated through English to introduce variations in the dataset. Additional augmentation has been performed using the ContextualWordEmbsAug class of the nlpaug library, which optimizes models like bert-base-uncased for word substitutions based on context. Underrepresented labels have been identified, and new samples have been generated to balance the dataset by applying augmentation to randomly selected rows. We have aimed to expand the size of the dataset and thus our training dataset has increased from 800 entries to 1600 for Malayalam and 808 entries to 1616 for Tamil.

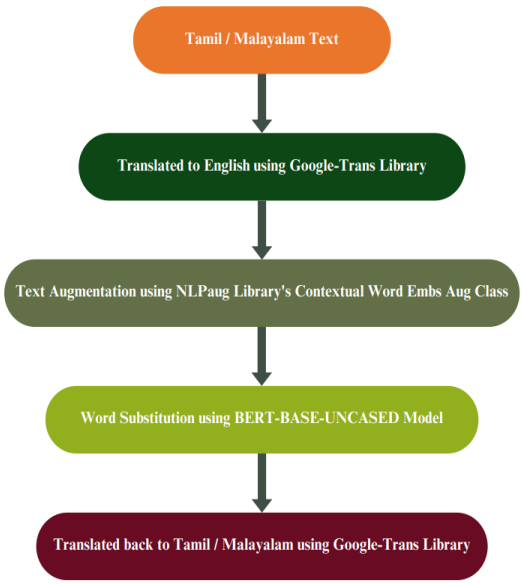


Figure 2: Data Augmentation technique

5 Results and Analysis

Our performance comparison among the ML, DL, and transformer-based approaches is shown in this section.

5.1 Parameter Setting

In Table 3, lr, optim, bs, wd, and wr represent learning_rate, optimizer, batch_size, weight_decay, and warmup_ratio respectively. Model name AHD represents AI-Human-Detector, E5-SLAGD represents E5–Small-Lora-Ai-Generated-Detector, and DAT represents Detect-Ai-Text.

Table 3: Parameter settings for different models

Model	lr	optim	bs	wd	wr
AHD	$2e^{-5}$	AdamW	4	0.01	0.1
e5-SLAGD	$2e^{-5}$	AdamW	4	0.01	0.1
DAT	$2e^{-5}$	AdamW	4	0.01	0.1
LSTM	$1e^{-3}$	Adam	32	-	-
BiLSTM	$1e^{-3}$	Adam	32	-	-
GRU	$1e^{-3}$	Adam	32	-	-
RNN	$1e^{-3}$	Adam	32	-	-

5.2 Evaluation Metrics

The performance of various models has been evaluated by calculating the precision (P), recall (R), and F1-Score on the test set.

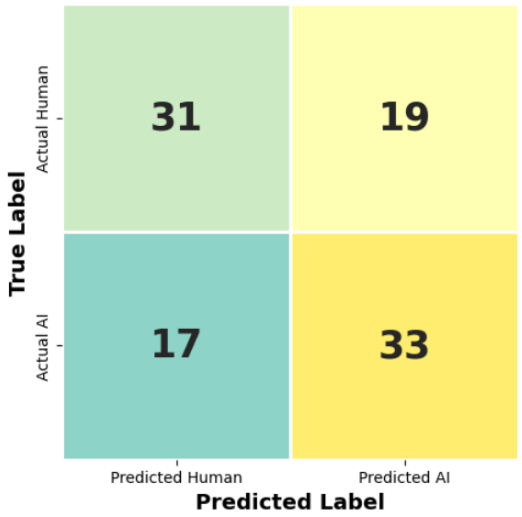
5.3 Comparative Analysis

Table 4: Performance of different systems on Malayalam and Tamil test datasets

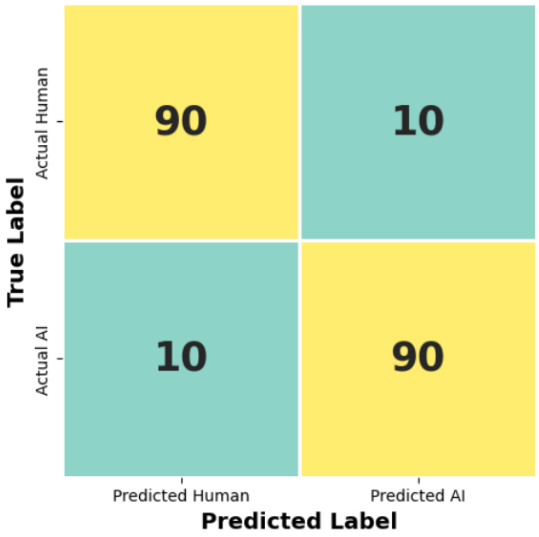
Classifier	Malayalam			Tamil		
	P	R	F1	P	R	F1
ML						
DT	0.15	0.09	0.11	0.15	0.09	0.13
RF	0.36	0.23	0.39	0.16	0.13	0.43
SVM	0.21	0.13	0.23	0.16	0.11	0.12
XGBOOST	0.32	0.11	0.35	0.21	0.13	0.35
DL						
BiLSTM	0.25	0.25	0.33	0.23	0.48	0.31
LSTM	0.25	0.50	0.33	0.23	0.48	0.31
GRU	0.77	0.53	0.49	0.46	0.51	0.38
RNN	0.43	0.43	0.37	0.27	0.52	0.36
TF						
E5-SLAGD	**0.91**	**0.90**	**0.90**	**0.65**	**0.62**	**0.59**
AHD	0.25	0.50	0.33	0.70	0.70	0.48
DAT	0.87	0.87	0.87	0.26	0.48	0.33

We have conducted a comparative study among four ML models Random Forest, Support Vector Machine, Decision Tree, and XGBoost, four DL models (LSTM, BiLSTM, GRU, and RNN), and three transformer-based models (AI-Human-Detector, Detect-AI-Text, and E5-Small-LoRA-AI-Generated-Detector). Among the ML models, XGBoost achieved the highest F1 score. DL model GRU has outperformed ML models. Transformer models have significantly outperformed both ML and DL models. The E5-Small-LoRA-AI-Generated-Detector (E5-SLAGD) model, with an F1 score of 0.8994 for Malayalam and 0.592 for Tamil, has achieved the best results ranking 7th in Malayalam and 28th in Tamil among all participants in the shared task.

5.4 Error Analysis



(a) Confusion Matrix for Malayalam



(b) Confusion Matrix for Tamil

Figure 3: Confusion Matrices for Malayalam and Tamil

Table 4 shows that E5-Small-Lora-AI-Generated-Detector did better than all others. We have created confusion matrices shown in Figure 3a and 3b) for a better understanding of the system. The False Positive (FP) for Malayalam was 4%, while FP for Tamil was 5% for E5.

AI-generated reviews often mimic human speech patterns, making misclassification a central issue. Tamil is morphologically richer than Malayalam and the overlapping of tokens is more.

Thus, the model faced greater challenges with Tamil compared to Malayalam.

6 Limitation

While NLTKTokenizer provided a baseline, our future work will investigate subword tokenization techniques such as Byte-Pair Encoding (BPE) or WordPiece, to better handle morphologically complex words. Also, the specific hyperparameter values (learning rate, optimizer, batch size, weight decay, and warmup ratio) were initially selected based on recommendations from the original model publications. Due to time constraints, a formal hyperparameter optimization strategy, such as grid search or Bayesian optimization, could not be employed. Our future works will explore more systematic hyperparameter tuning methods to potentially improve model performance.

7 Conclusion

In this paper, a comparative study has been carried out among various machine learning, deep learning, and transformer-based models for detecting AI-generated product reviews in Malayalam and Tamil languages. We have utilized the dataset of the "Shared Task on Detecting AI-Generated Product Reviews in Dravidian Languages" at NAACL for training the models. Three pre-trained models were used among which, AI-Human-Detector has outperformed other models with an F1 score of 0.8994 for Malayalam and 0.5877 for Tamil. We have calculated the Jaccard Index for both languages to quantify token overlap between AI-generated and human-written reviews. Our average Jaccard Index for Tamil at 0.35 is greater than that for Malayalam at 0.22. This suggests the token overlap in Tamil is higher where AI generated and human written have a higher percentage of tokens. It is likely that this larger overlap is why the two classes are more difficult to distinguish in Tamil. The overlap between the target tokens in the dataset has caused misclassification despite the Transformer-based models having excellent contextual understanding. The tokenization of Tamil is more complicated due to richer morphology for which the Tamil model performed worse than the Malayalam model. To tackle these issues in the near future, we plan to explore advanced transformer architectures and enhanced data augmentation.

Ethical Statement

The data analysis and model development tools and technologies used for this study are used ethically and responsibly. The purpose of our work is to create a system that detects AI product reviews for the good of maintaining transparency and authenticity on online platforms. We believe knowledge is for sharing, so we will share our work and contribute to the development of AI-generated content detection in low-resource languages such as Malayalam and Tamil.

References

Harika Abburi, Kalyani Roy, Michael Suesserman, Nirmala Pudota, Balaji Veeramani, Edward Bowen, and Sanmitra Bhattacharya. 2023. A simple yet efficient ensemble approach for ai-generated text detection. *Preprint*, arXiv:2311.03084.

Hamed Alshammari, Ahmed El-Sayed, and Khaled Elleithy. 2024. Ai-generated text detector for arabic language using encoder-based transformer architecture. *Big Data and Cognitive Computing*, 8(3).

Ilker Cingillioglu. 2023. Detecting ai-generated essays: the chatgpt challenge. *The International Journal of Information and Learning Technology*, 40(3):259–268.

Raghav Gaggar, Ashish Bhagchandani, and Harsh Oza. 2023. Machine-generated text detection using deep learning. *Preprint*, arXiv:2311.15425.

Niful Islam, Debopom Sutradhar, Humaira Noor, Jarin Tasnim Raya, Monowara Tabassum Maisha, and Dewan Md Farid. 2023. Distinguishing human generated text from chatgpt generated text using machine learning. *Preprint*, arXiv:2306.01761.

Narek Maloyan, Bulat Nutfullin, and Eugene Ilyushin. 2022. Dialog-22 ruatd generated text detection. *ArXiv*, abs/2206.08029.

Teodor-George Marchitan, Claudiu Creanga, and Liviu P. Dinu. 2024. Transformer and hybrid deep learning based models for machine-generated text detection. *Preprint*, arXiv:2405.17964.

George K. Mikros, Athanasios Koursaris, Dimitrios Bilianos, and George Markopoulos. 2023. Ai-writing detection using an ensemble of transformers and stylometric features. In *IberLEF@SEPLN*.

Yuhong Mo, Hao Qin, Yushan Dong, Ziyi Zhu, and Zhenglin Li. 2024. Large language model (llm) ai text generation detection based on transformer deep learning algorithm. *ArXiv*, abs/2405.06652.

Duke Nguyen, Khaing Myat Noe Naing, and Aditya Joshi. 2023. Stacking the odds: Transformer-based

ensemble for ai-generated text detection. pages 173–178.

B. Pavlyshenko and Mykola Stasiuk. 2023. Augmentation in a binary text classification task. *2023 IEEE 13th International Conference on Electronics and Information Technologies (ELIT)*, pages 177–180.

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.

Andric Valdez-Valenzuela, Ricardo Loth Zavala-Reyes, Victor Giovanni Morales Murillo, and Helena Gómez-Adorno. 2024. The iimasnlp team at iberautextification 2024: Integrating graph neural networks, multilingual llms, and stylometry for automatic text identification.