AnalysisArchitects@DravidianLangTech 2025: BERT Based Approach For Detecting AI Generated Product Reviews In Dravidian Languages

Abirami Jayaraman abirami2210382@ssn.edu.in
Aruna Devi Shanmugam aruna2210499@ssn.edu.in
Dharunika Sasikumar dharunika2210459@ssn.edu.in
Bharathi B bharathib@ssn.edu.in

Department of Computer Science and Engineering
Sri Sivasubramaniya Nadar College of Engineering
Kalavakkam, Chennai, Tamil Nadu

Abstract

The shared task on Detecting AI-generated Product Reviews in Dravidian Languages is aimed at addressing the growing concern of AI-generated product reviews, specifically in Malayalam and Tamil. As AI tools become more advanced, the ability to distinguish between human-written and AI-generated content has become increasingly crucial, especially in the domain of online reviews where authenticity is essential for consumer decision-making. In our approach, we used the ALBERT, IndicBERT, and Support Vector Machine (SVM) models to classify the reviews. The results of our experiments demonstrate the effectiveness of our methods in detecting AI-generated content.

1 Introduction

The proliferation of AI-generated content has raised significant concerns across various domains, particularly in online reviews where authenticity is paramount for consumer decision-making. The Shared Task on Detecting AI-generated Product Reviews in Dravidian Languages(Premjith et al., 2025) addresses this issue by focusing on the detection of AI-generated reviews in Malayalam and Tamil. As AI tools become more sophisticated, distinguishing between human-written and AI-generated content has become increasingly challenging and crucial.

Online reviews play a critical role in influencing consumer behavior and purchasing decisions. However, the rise of generative AI has led to an increase in fake reviews, which can undermine consumer trust and distort market dynamics. Luo et al.(Luo et al., 2023) highlight the impact of AI-generated fake reviews on e-commerce platforms and propose a supervised learning approach to detect such reviews. Their study emphasizes the importance of developing robust detection methods to maintain the integrity of online reviews.

In this shared task, datasets containing both human-written and AI-generated reviews in Malayalam and Tamil were provided. The objective was to develop models capable of accurately classifying these reviews while addressing the specific challenges posed by Dravidian languages. Models including ALBERT, IndicBERT, and Support Vector Machine (SVM) classifiers were utilized in this approach.

The effectiveness of these models in detecting AI-generated reviews in Malayalam and Tamil was demonstrated through our experiments. Section 2 of this paper provides a brief summary about various works done in this filed. Section 3 provides the description of the datasets. Section 4 explains the methodology used. Section 5 provides detailed information about the implementation of the methodology. Section 6 provides a consolidated view of the results obtained in the test. Section 7 elaborates the shortcomings of these models. Section 8 concludes the paper.

2 Related Works

The study by Gupta and Jindal (Gupta et al., 2024) highlights the challenge of AI-generated fake reviews, which are produced using generative AI tools, complicating the integrity of online feedback systems. To combat this issue, various detection techniques are employed, including rule-based approaches that utilize predefined characteristics of fake reviews, graph-based techniques that analyze user-review relationships for anomalies, machine learning algorithms trained on review features to classify authenticity, and deep learning models that capture complex patterns in the data. The research emphasizes ongoing efforts to enhance detection accuracy and identifies key challenges, such as the

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 661–665
May 3, 2025 ©2025 Association for Computational Linguistics

evolving tactics of those generating fake reviews, particularly through sophisticated AI-generated content.

This study by Jean Michel Sahut (Sahut et al., 2024) presents a novel supervised learning approach aimed at distinguishing between human-written reviews and those generated by AI. The study constructs various variables and employs an outlier detection method based on cumulative probability density to enhance detection accuracy. It demonstrates that the proposed method outperforms existing baseline techniques in identifying AI-generated reviews.

The study by Mudasir Ahmad Wani et al (Wani et al., 2024) introduces a framework utilizing deep learning algorithms and natural language processing (NLP) techniques to detect AI-generated spam reviews. The framework integrates multiple deep learning architectures, such as CNNs and LSTMs, and applies advanced NLP methods for thorough textual analysis, proving effective across diverse datasets.

The study by Jiwei et al (Mohammed and Ahmed, 2023) examines the impact of AI-generated reviews on consumer perceptions in online shopping. It discusses how AI tools can manipulate buyer behavior by producing reviews that may misrepresent products. While these reviews can enhance the shopping experience by summarizing key features, they also raise concerns about reliability and trust in e-commerce. The study emphasizes the need for vigilance regarding the integration of AI-generated content in online platforms, as it significantly influences market dynamics and consumer trust.

The paper by Anna Shcherbiak et al(Shcherbiak et al., 2024) investigates the effectiveness of various classifiers in distinguishing between texts generated by AI and those written by humans. The study employs a dataset comprising both AI-generated and human-written texts, utilizing advanced machine learning models to perform the classification task.

The research by Lorenz Mindner et al(Mindner et al., 2023) explores various features to detect AI-generated texts, including those rephrased by AI. The study uses a new text corpus and achieves high F1-scores in classifying both basic and advanced human-generated and AI-generated texts

Labels	Malayalam	Tamil
AI	406	401
HUMAN	394	408

Table 1: Split up of training data into 2 classes for Malayalam an Tamil

Labels	Malayalam	Tamil
AI	100	48
HUMAN	100	52

Table 2: Split up of testing data into 2 classes for Malayalam an Tamil

3 Dataset Description

The dataset used in this study consists of Tamil and Malayalam reviews, which include both human-written and AI-generated text. It is divided into training and validation sets to help with effective classification.

For the Malayalam dataset, there are 801 training samples, categorized into two labels: Human-written and AI-generated and are tabulated in Table 1. The test set contains 201 samples, which are used to evaluate the performance of the model. A detailed breakdown of this dataset can be seen in Table 2.

Similarly, the Tamil dataset has 809 training samples, also classified under the same two labels and are tabulated in Table 1. The test set consists of 101 samples, which help assess the accuracy of classification models. The distribution of this dataset is provided in Table 2.

4 Proposed Work

The goal of this paper is to classify Tamil and Malayalam text data to determine whether the content is human or AI-generated. Advanced transformer-based language models are employed for this task due to their efficiency and strong performance. The dataset, consisting of Tamil and Malayalam text samples with labels indicating "HUMAN" or "AI," is first preprocessed. Labels are encoded numerically, with "HUMAN" mapped

to 0 and "AI" to 1.

In this study, the three models, ALBERT(Lan et al., 2020), IndicBERT(Kakwani et al., 2020), and SVM, were used. After training, the model was evaluated on the test data to generate predictions. Metrics such as accuracy and macro F1-score were used to assess the model's ability to distinguish between human and AI-generated text.

5 Experimental Results

Implementation involves using ALBERT (A Light Bidirectional Encoder Representations from Transformers), Indic-BERT, and SVM (Support Vector Machine) models to classify Tamil and Malayalam text data as human-generated or AI-generated. The data preprocessing steps are largely similar for all three models, with differences arising primarily during the training process.

Labels are converted into numeric values using the label encoding. Specifically, the label "HUMAN" is encoded as 0, and "AI" is encoded as 1. This encoding step ensures that the data are compatible with machine learning models, which require numeric labels for classification tasks. After label encoding, the data is split into training and validation sets. Typically, 75% of the data is allocated for training, and the remaining 25% is reserved for validation. Text data is then tokenized using appropriate tokenizers for each model. Tokenization is the process of converting the text into smaller units (tokens), padding or truncating the sequences to a fixed length, and converting them into numerical representations. This ensures that the text data is in a format suitable for the respective models.

ALBERT model is a transformer-based architecture designed for efficient text classification tasks. The model and its tokenizer are loaded using the Hugging Face Transformers library. The tokenizer preprocesses the text data by converting it into a format compatible with ALBERT. This involves tokenizing the text, padding or truncating sequences to a fixed length, and converting the tokens into numerical values. The training process is configured using the Training Arguments class, where hyperparameters such as learning rate, batch size, number of epochs, and weight decay are specified.

Model used	Accuracy	Macro F-1 score
ALBERT	0.9136	0.9122
IndicBERT	0.9653	0.9651
SVM	0.8465	0.8465

Table 3: Performance on Tamil training dataset

Model used	Accuracy	Macro F-1 score
ALBERT	0.950	0.9499
IndicBERT	0.965	0.9649
SVM	0.775	0.7748

Table 4: Performance on Malayalam training dataset

Model is trained using the Trainer class, which handles the training loop, including backpropagation and optimization. Once the model is trained, it is evaluated on the validation set. The test data is tokenized similarly to the training data and passed through the trained ALBERT model to obtain predictions. These predictions are mapped back to their original labels ("HUMAN" or "AI") for interpretability.

The Indic-BERT model is another transformer-based model, specifically designed for Indic languages like Tamil and Malayalam. The training process for Indic-BERT follows a similar approach to ALBERT. The model and tokenizer are loaded from the Hugging Face library, and the text data is tokenized into a format suitable for the model. The training configuration is set up using the Training Arguments class, and the model is trained using the Trainer class, similar to the ALBERT approach. The evaluation process is the same as with ALBERT, where the model is tested on the validation set, and the predictions are mapped back to their original labels.

For the SVM model, the process differs from the transformer-based models. Instead of tokenizing the text data into tokens and using pre-trained models, SVM requires feature extraction. The text data is converted into numerical features using techniques such as TF-IDF (Term Frequency-Inverse Document Frequency). These features are then used to train the SVM classifier.

The SVM model is trained with linear kernel. The training process involves finding the optimal hyperplane that separates the human and

Model used	Accuracy	Macro F-1 score
ALBERT	0.46	0.4375
IndicBERT	0.62	0.62
SVM	0.66	0.6594

Table 5: Performance on Tamil testing dataset

Model used	Accuracy	Macro F-1 score
ALBERT	0.885	0.8849
IndicBERT	0.5	0.333
SVM	0.68	0.6797

Table 6: Performance on Malayalam testing dataset

AI-generated text. The performance of the SVM model is evaluated using metrics such as accuracy and macro F1-score. Once trained, the SVM model is used to make predictions on the validation set. The predictions are then mapped back to the original labels ("HUMAN" or "AI").

The evaluation metrics used include accuracy and macro F1-score. Accuracy measures the overall correctness of the predictions, while the macro F1-score provides a balanced evaluation by considering both precision and recall for each class and averaging them. The performance of these models on Tamil training dataset is tabuated in Table 3 and on Malayalam training set is tabulated in Table 4

For all models, the results are saved to a TSV file, which includes the ID of each test instance and its corresponding predicted label. These results are then analyzed to assess the performance of each model in distinguishing between human and AI-generated text. Implementation of all these models are available in Github.[1]

6 Result

The performance of all the models on the Tamil testing dataset has been listed in Table 5, and on the Malayalam testing dataset in Table 6. From these tabulations, it can be inferred that better performance was achieved by SVM on the Tamil dataset and it's confusion matrix is available in Figure 1, while ALBERT performed better on the Malayalam dataset and it's confusion matrix is available in Figure 2. Although IndicBERT

[1]Github page : https://github.com/S-ArunaDevi06/AI_generated_review_detection/

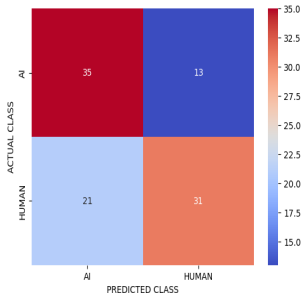

Figure 1: Confusion matrix for performance of SVM on Tamil testing dataset

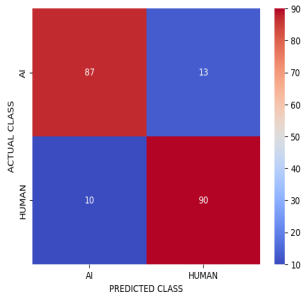

Figure 2: Confusion matrix for performance of ALBERT on Malayalam testing dataset

performed well on the training dataset for both languages, its performance on the testing dataset was comparatively lower for both languages.

By analysing the performance of SVM on tamil testing dataset, AI-written sentences that are misclassified as human-written sentences often contain structured, well-articulated language. By analysing the performance of ALBERT on Malayalam testing dataset, it can be observed that the model often struggles with sentences that have an informal or conversational tone. This model also struggles with sentences have transliterated words.

7 Limitations

In the ALBERT model, the maximum token length (512) can be restrictive for longer AI-generated content. If the dataset contains AI text from only one model (e.g., ChatGPT, GPT-3), ALBERT may overfit and fail on texts from newer AI models (e.g., Gemini, Claude).

Like ALBERT, IndicBERT might struggle to detect AI-generated text from newer AI models. SVM treats text as vectors without sequential information, it struggles with coherence and fluency patterns that distinguish AI from human

text. SVM works well for clearly separable classes. But if AI-generated and human text are very similar, SVM fails.

8 Conclusions

In this shared task, various BERT models as well as SVM were evaluated for classification tasks on Dravidian languages. The 23rd place was secured using SVM for Tamil, and the 9th place was secured using ALBERT for Malayalam in this shared task.

References

Richa Gupta, Vinita Jindal, and Indu Kashyap. 2024. Recent state-of-the-art of fake review detection: a comprehensive review. *The Knowledge Engineering Review*, 39:e8.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., 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, Online. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. *Preprint*, arXiv:1909.11942.

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

Lorenz Mindner, Tim Schlippe, and Kristina Schaaff. 2023. *Classification of Human- and AI-Generated Texts: Investigating Features for ChatGPT*, page 152–170. Springer Nature Singapore.

Ennaouri Mohammed and Zellou Ahmed. 2023. Machine learning approaches for fake reviews detection: A systematic literature review. *Journal of Web Engineering*.

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.

Jean Michel Sahut, Michel Laroche, and Eric Braune. 2024. Antecedents and consequences of fake reviews in a marketing approach: An overview and synthesis. *Journal of Business Research*, 175:114572.

Anna Shcherbiak, Hooman Habibnia, Robert Böhm, and Susann Fiedler. 2024. Evaluating science: A comparison of human and ai reviewers. *Judgment and Decision Making*, 19:e21.

Mudasir Ahmad Wani, Mohammed ElAffendi, and Kashish Ara Shakil. 2024. Ai-generated spam review detection framework with deep learning algorithms and natural language processing. *Computers*, 13(10).