SSNTrio@DravidianLangTech 2025: Identification of AI Generated Content in Dravidian Languages using Transformers

Bhuvana J
Sri Sivasubramaniya Nadar College of Engineering
bhuvanaj@ssn.edu.in

Mirnalinee T T
Sri Sivasubramaniya Nadar College of Engineering
MirnalineeTT@ssn.edu.in

Rohan R
Sri Sivasubramaniya Nadar College of Engineering
rohan2210124@ssn.edu.in

Diya Seshan
Sri Sivasubramaniya Nadar College of Engineering
diya2210208@ssn.edu.in

Avaneesh Koushik
Sri Sivasubramaniya Nadar College of Engineering
avaneesh2210179@ssn.edu.in

Abstract

The increasing prevalence of AI-generated content has raised concerns about the authenticity and reliability of online reviews, particularly in resource-limited languages like Tamil and Malayalam. This paper presents an approach to the Shared Task on Detecting AI-generated Product Reviews in Dravidian Languages at NAACL2025, which focuses on distinguishing AI-generated reviews from human-written ones in Tamil and Malayalam. Several transformer-based models, including IndicBERT, RoBERTa, mBERT, and XLM-R, were evaluated, with language-specific BERT models for Tamil and Malayalam demonstrating the best performance. The chosen methodologies were evaluated using Macro Average F1 score. In the rank list released by the organizers, team SSNTrio, achieved ranks of 3rd and 29th for the Malayalam and Tamil datasets with Macro Average F1 Scores of 0.914 and 0.598 respectively.

1 Introduction

The swift progress in AI-generated text has sparked apprehensions regarding authenticity and dependability across multiple sectors, particularly in online reviews, where user-generated content plays a crucial role in shaping consumer choices. As AI technologies evolve, it has become increasingly difficult to differentiate between reviews authored by humans and those produced by AI. To maintain transparency on online platforms, it is essential to implement effective methods for detecting AI-generated content, especially in low-resource languages such as Tamil and Malayalam, which are characterized by a scarcity of annotated datasets and linguistic tools.

Identifying AI-generated content in Tamil and Malayalam poses distinct challenges due to their complex morphology, agglutinative structure, and varied writing styles. Standard multilingual models frequently struggle to grasp the linguistic nuances of these languages, highlighting the need for tailored approaches. Transformer based models, particularly those fine-tuned for Tamil and Malayalam, have demonstrated potential in overcoming these obstacles by developing context-aware representations suited to these languages.

This paper outlines a methodology for the Shared Task (Premjith et al., 2025) on Detecting AI-generated Product Reviews in Dravidian Languages, which centers on the classification of AI-generated versus human-written reviews in Tamil and Malayalam. The objective of the task was to assess the efficiency of various models in recognizing AI-generated text and to investigate the challenges that are specific to Dravidian languages.

2 Related Works

The detection of AI-generated text has become a critical area of research due to the rise of generative AI tools such as ChatGPT. Several studies (Mindner et al., 2023) have aimed to distinguish between human-written and AI-generated content, employing a range of features such as perplexity, semantic analysis, readability, and AI feedback. These features have been used to improve detection accuracy across various types of AI-generated texts. BERT-based models (Javaji et al., 2024), for example, have demonstrated impressive performance, achieving F1-scores over 96% for identifying basic AI-generated texts and over 78% for AI-rephrased texts. These models are particularly effective due to their ability to capture contextual and semantic nuances within text.

Other studies (Mozes et al., 2021) have explored the issue of adversarial attacks on natural language processing (NLP) models, revealing that humans are capable of generating adversarial examples through semantic-preserving word substitutions. These attacks have raised important questions about the vulnerability of NLP systems and the importance of validating adversarial inputs for

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 413–417

maintaining text integrity. Moreover, the identification of AI-generated content extends to broader concerns such as plagiarism, fake news, and content authenticity. For instance, large datasets consisting of essays have been employed to train machine learning models that classify content as either AI-generated or human-written. This approach (Wang et al., 2024) is particularly relevant in academic and online environments where ensuring the authenticity of text is paramount.

Additionally, efforts (Abburi et al., 2023) have been made to attribute specific language models to AI-generated text, as demonstrated by ensemble neural models that combine pre-trained LLMs to classify text. These advances reflect the growing importance of detecting AI-generated material in various domains, from education to online content moderation, and emphasize the need for robust and efficient detection systems.

It is clear that, Language-specific BERT models have not been utilized for this task. Leveraging the potential of such models by using their advanced architectures could serve as a novel contribution to this study.

3 Dataset Description

This task is focused on the precise identification of AI-generated reviews in Dravidian languages. It consists of two subtasks, differentiated by language: Tamil and Malayalam. The training dataset, as provided by the organisers, comprises the Review ID, the corresponding review text, and the associated label indicating whether the review is generated by AI or Human. Refer to Table 1 for detailed dataset statistics of each language.

	Tamil	Malayalam
Training Data	808	800
Testing Data	100	200

Table 1: Dataset Statistics

Figure 1 and Figure 2 show the distribution of labels in the Malayalam and Tamil datasets, respectively.

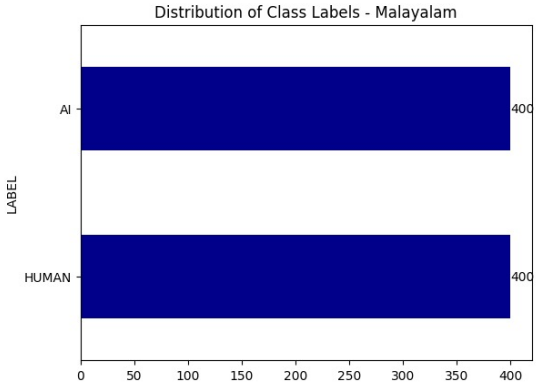

Figure 1: Label Distribution in Malayalam Dataset

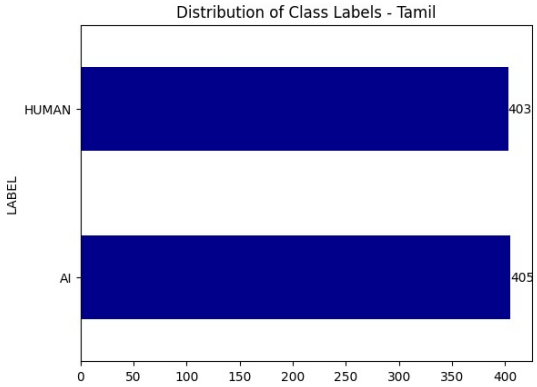

Figure 2: Label Distribution in Tamil Dataset

4 Methodology

4.1 Data Preprocessing

The methodology starts with preprocessing the dataset to ensure consistency and prepare the text for model training. Text data is converted to lowercase to maintain uniformity. Punctuation is removed to reduce noise, and the text is tokenized into individual words. Stopwords are removed to focus on meaningful terms, enhancing the quality of the textual data. The cleaned text is then reassembled for downstream processing. Additionally, the labels "AI" and "Human" are encoded to facilitate the model's understanding and classification tasks.

4.2 Synthetic Sample Generation

Recognizing the challenge of a limited dataset, SMOTE (Synthetic Minority Over-sampling Technique) upsampling was used to generate synthetic samples, enriching the dataset. This technique creates new synthetic examples by interpolating between existing samples, helping to enhance data diversity while preserving meaningful patterns. SMOTE mitigates class imbalance by ensuring that the classifier is exposed to a more balanced distribution of examples, reducing bias towards the majority class. Through this method, the dataset size was increased to 1,200 samples per label, improving the model's ability to generalize across all label categories and enhancing classification per-

formance.

4.3 Tokenisation and Feature Representation

For feature representation, the preprocessed text undergoes tokenization with the BERT tokenizer. The tokenizer transforms text input into numerical sequences, consisting of input IDs and attention masks, ensuring it works with the pre-trained BERT model. For consistency, padding and truncation are utilized, restricting the tokenized sequences to a maximum length of 128 tokens. This approach guarantees that the model can efficiently manage textual inputs of different lengths.

4.4 Model Building

After preprocessing and tokenization, the dataset was split into training and testing sets. The tokenizer was fit on the training data and applied to the test data to maintain consistency. Various pre-trained transformer models, including BART, IndicBERT, and RoBERTa, were tested, but Tamil BERT and Malayalam BERT performed best. These models likely excelled due to their language specialization, improving accuracy and contextual understanding.

The fine-tuned Tamil and Malayalam BERT models (Joshi, 2022) were trained on the respective datasets for 25 epochs, using the Adam optimizer (Zhang, 2018) and sparse categorical cross-entropy as the loss function. Both models were designed to handle binary classification tasks efficiently, leveraging their contextual embeddings to predict target labels accurately. This approach ensured robust performance while maintaining the interpretability and scalability of the solution.

5 Performance Metrics

This section provides insight on the metrics used to evaluate the performance of the methodologies employed for each task.

- **Recall** for a specific label indicates the proportion of correctly identified instances of that label out of all true instances in the data.
- **Precision** for a specific label measures the proportion of correctly identified instances of that type out of all instances predicted as that type by the model.
- **F1 Score** for a specific label is the harmonic mean of precision and recall for that label, providing a balanced measure.
- **Macro-Average F1 Score** evaluates the model's performance across all classes in multi-label classification by calculating the F1 score for each class independently, ensuring that each class is given equal weight in the overall score.

Language	F1 Score	Precision	Recall	Accuracy
Malayalam	0.91	0.92	0.92	0.92
Tamil	0.59	0.73	0.65	0.66

Table 2: Performance Scores

6 Result Analysis

Several transformer models were evaluated, including IndicBERT, RoBERTa, mBERT, XLM-R (Conneau et al., 2019), and others, for the task of distinguishing AI-generated content from human-written text in Tamil and Malayalam. The Malayalam and Tamil BERT models consistently demonstrated superior performance, proving to be the most effective for text classification in these languages.

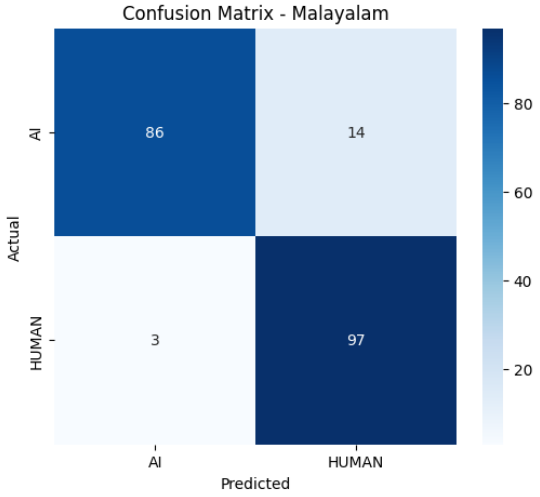

Figure 3: Confusion Matrix - Malayalam

Additionally, it was noticed that SMOTE upsampling played a crucial role in improving model performance by increasing the dataset size. This enhancement significantly boosted the model's ability to generalize across different categories, leading to a significant improvement in classification accuracy and overall robustness.

Using this model, test datasets were classified and key evaluation metrics were computed, including accuracy, precision, recall, and F1 scores for each language. The final performance results are presented in Table 2. The submissions ranked 3rd

for Malayalam and 29th for Tamil in the official ranklist released by the organizers.

The confusion matrices shown in Figures 3 and 4 depict the model's classification accuracy and the distribution of predicted labels compared to the actual labels for the Malayalam and Tamil test sets, respectively.

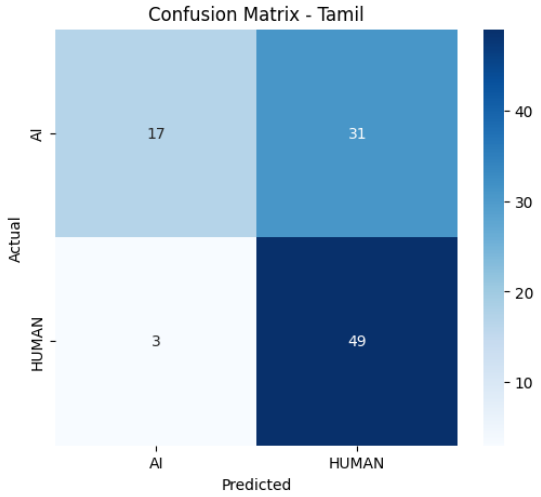

Figure 4: Confusion Matrix - Tamil

The model demonstrated high recall for Human text, indicating strong recognition of human-written content. However, there were challenges in correctly identifying AI text. This suggests that there may be feature overlap between AI and Human texts, and further model refinement is needed to improve AI content detection.

7 Conclusion

In conclusion, this paper has investigated the identification of AI-generated reviews for Indian languages, focusing mainly on Tamil and Malayalam.

The Malayalam and Tamil BERT models demonstrated superior performance for several reasons. First, these language-specific BERT models were fine-tuned on large corpora of Tamil and Malayalam texts, enabling them to better capture the unique syntactic and semantic nuances of these languages.

Second, the models' architecture was better suited to handle the linguistic characteristics of Tamil and Malayalam, which are significantly different from other languages, resulting in improved classification accuracy. This fine-tuning contributed to more precise identification of AI-generated content compared to other models.

The evaluation and analysis of the results from the tasks provided valuable insights into the challenges of Identifying and classification in multilingual contexts. Ongoing advancements in model refinement and feature extraction will be crucial for enhancing performance in future research efforts.

8 Future Enhancements

Neural networks can enhance AI-generated text detection by leveraging deep learning architectures. Transformer-based models like BERT, RoBERTa, and XLM-R can be fine-tuned for Tamil and Malayalam to improve classification accuracy. Hybrid models combining CNNs for feature extraction and LSTMs for sequential learning can further enhance performance. Additionally, explainable AI techniques like LIME and SHAP can provide insights into model decisions. Future work can also explore adversarial training and semi-supervised learning to improve robustness and generalization. Incorporating contrastive learning can help the model better distinguish between subtle differences in AI and human-generated text. Moreover, multi-modal approaches that integrate speech and text analysis could further improve detection accuracy in social media and real-world applications.

9 Limitations

Despite strong performance, the Tamil and Malayalam BERT models face limitations. Their effectiveness depends on the quality and balance of training data, with biases affecting classification accuracy. Additionally, the dataset may not fully represent real-world reviews, limiting generalization across domains.

A key challenge is the presence of domain-specific biases. Models trained on limited sources may struggle with sectors like healthcare or finance. Code-mixing and transliteration in Tamil and Malayalam further complicate classification, leading to errors.

Furthermore, the model's applicability to other Indian languages remains uncertain without fine-tuning. An important area for further exploration is error analysis, which could help identify issues related to ambiguous structures, data limitations, or model biases. Finally, as AI-generated content evolves to mimic human writing more closely, maintaining high classification accuracy will require periodic retraining with updated datasets and the incorporation of more advanced linguistic features.

References

Harika Abburi, Michael Suesserman, Nirmala Pudota, Balaji Veeramani, Edward Bowen, and Sanmitra Bhattacharya. 2023. Generative ai text classification using ensemble llm approaches. *Preprint*, arXiv:2309.07755.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Prashanth Javaji, Pulaparthi Satya Sreeya, and Sudha Rajesh. 2024. Detection of ai generated text with bert model. In *2024 2nd World Conference on Communication Computing (WCONF)*, pages 1–6.

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

Lorenz Mindner, Tim Schlippe, and Kristina Schaaff. 2023. Classification of human- and ai-generated texts: Investigating features for chatgpt. In *Artificial Intelligence in Education Technologies: New Development and Innovative Practices*, pages 152–170, Singapore. Springer Nature Singapore.

Maximilian Mozes, Max Bartolo, Pontus Stenetorp, Bennett Kleinberg, and Lewis D. Griffin. 2021. Contrasting human- and machine-generated word-level adversarial examples for text classification. *CoRR*, abs/2109.04385.

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.

Hao Wang, Jianwei Li, and Zhengyu Li. 2024. Ai-generated text detection and classification based on bert deep learning algorithm. *Preprint*, arXiv:2405.16422.

Zijun Zhang. 2018. Improved adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*, pages 1–2. Ieee.