InnovateX@DravidianLangTech 2025: Detecting AI-Generated Product Reviews in Dravidian Languages

Moogambigai A, Pandiarajan D, Bharathi B
Department of Computer Science and Engineering
Sri Sivasubramaniya Nadar College of Engineering, Chennai, India
moogambigai2370071@ssn.edu.in
pandiarajan2370062@ssn.edu.in
bharathib@ssn.edu.in

Abstract

This paper presents our approach to the Shared Task on Detecting AI-Generated Product Reviews in Dravidian Languages as part of DravidianLangTech@NAACL 2025, as described by (Premjith et al., 2025). The task focuses on distinguishing between human-written and AI-generated reviews in Tamil and Malayalam, languages rich in linguistic complexities. Using the provided datasets, we implemented machine learning and deep learning models, including Logistic Regression (LR), Support Vector Machine (SVM), and BERT. Through preprocessing techniques like tokenization and TF-IDF vectorization, we achieved competitive results, with our SVM and BERT models demonstrating superior performance in Tamil and Malayalam respectively. Our findings underscore the unique challenges of working with Dravidian languages in this domain and highlight the importance of robust feature extraction.

1 Introduction

The proliferation of AI-generated content has brought both opportunities and challenges across various domains, including e-commerce, social media, and journalism. While AI can generate text efficiently, it also raises significant concerns regarding authenticity, particularly in online reviews (Kovács, 2024), where fake or AI-generated content can manipulate consumer trust and market dynamics. Detecting such content is essential for ensuring credibility and maintaining user trust in digital platforms (Diaz-Garcia and Carvalho, 2025).

This shared task focuses on detecting AI-generated product reviews in Dravidian languages, specifically Tamil and Malayalam (Priyadharshini et al., 2021). These languages pose unique challenges for natural language processing (NLP) due to their rich morphology, agglutinative nature, code-mixing tendencies, and lack of extensive annotated datasets. Tamil and Malayalam also frequently incorporate regional slang, making text analysis even more complex.

Our work addresses these challenges by employing both traditional machine learning methods, such as Support Vector Machine (SVM) and Logistic Regression (LR), and advanced deep learning approaches like BERT. We preprocess the data to capture linguistic nuances, leveraging techniques such as tokenization, stopword removal, and feature extraction using TF-IDF (Kumari et al., 2023). By comparing these models, we aim to identify systems that effectively distinguish human-written content from AI-generated text (Knight et al., 2023), while also contributing insights to the broader field of AI-generated content detection in low-resource languages.

Keywords

AI-generated content detection, Dravidian languages, Tamil and Malayalam, machine learning models, BERT, code-mixed text classification

2 Related Work

Detecting AI-generated content has been a growing area of research (Aho and Ullman, 1972), particularly in high-resource languages like English. Techniques such as transformer-based models (e.g., BERT) and traditional machine learning approaches (e.g., Support Vector Machines and Logistic Regression) have shown significant promise in identifying machine-generated text (Joshi et al., 2024). Studies utilizing BERT and its variants demonstrate strong performance in detecting patterns specific to AI-generated text (Shaik Vadla et al., 2024), leveraging contextual embeddings for improved classification accuracy. Traditional methods employing TF-IDF features combined with machine learning classifiers like SVM and Naive Bayes have also been effective in text classification

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 302–307

tasks, particularly in resource-constrained settings.

However, research on low-resource languages, such as Tamil and Malayalam, remains scarce despite their increasing presence in online spaces (American Psychological Association, 1983). Dravidian languages exhibit unique linguistic characteristics, including agglutination, rich morphology, and context-sensitive meaning, which make text processing challenging. Additionally, code-mixing and transliteration common in social media text add complexity to language modeling tasks (Abeera et al., 2023).

Prior work on Dravidian languages has primarily addressed sentiment analysis, sarcasm detection, and offensive language identification (Chandra et al., 1981). While these tasks share similarities with content classification, they do not specifically target the detection of AI-generated text (Ojo et al., 2024). Furthermore, the limited availability of annotated datasets and preprocessing tools tailored to Tamil and Malayalam constrains the applicability of standard NLP methods.

Building on these foundations (Abiola et al., 2025), this study investigates the applicability of both traditional machine learning models, such as Support Vector Machine (SVM) and Logistic Regression (LR), and deep learning approaches, including BERT and its multilingual variants, for detecting AI-generated product reviews in Tamil and Malayalam (Andrew and Gao, 2007). The research also considers the impact of linguistic characteristics such as code-mixing and transliteration (Rasooli and Tetreault, 2015), aiming to bridge the gap in AI-generated content detection for low-resource languages.

3 Dataset

The datasets provided in the shared task consisted of human-written and AI-generated product reviews in Tamil and Malayalam, structured with distinct features and labeled samples. The data distribution is presented in Table 1.

- **Training Data:** Tamil comprised 808 comments, while Malayalam contained 800, with features including ID, DATA, and LABELS.
- **Test Data:** Tamil included 100 comments, and Malayalam comprised 200, with features restricted to ID and DATA.

Comprehensive preprocessing was performed to standardize the datasets, including feature encoding, label normalization, and partitioning into training and evaluation subsets. This ensured optimal compatibility and performance across both traditional and transformer-based models (Javaji et al., 2024).

4 Methodology

Our approach to distinguishing human-written and AI-generated reviews in Tamil and Malayalam involved leveraging both traditional and transformer-based models. The methodology is detailed in the Figure 1:

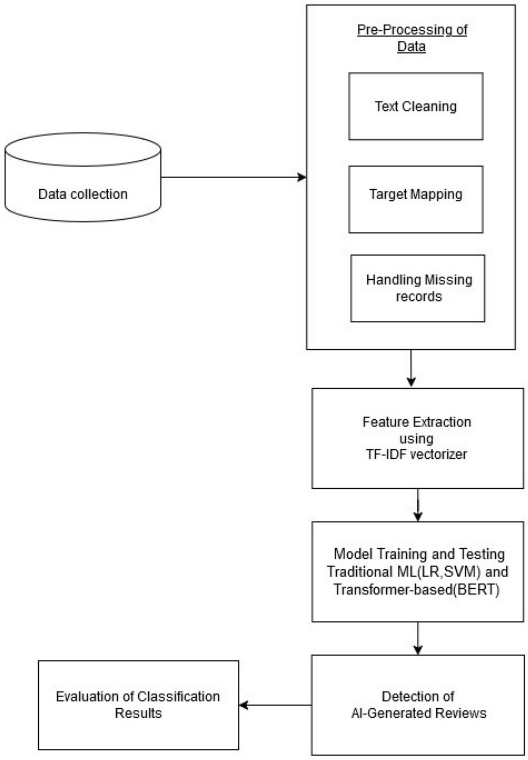

Figure 1: Framework of Proposed Methodology

4.1 Models for Classification

- **Logistic Regression (LR):** Utilized TF-IDF vectorization to transform textual data into numerical features, enabling efficient linear classification (Bhargav and Dhanalakshmi, 2024).
- **Support Vector Machine (SVM):** Combined robust preprocessing techniques with traditional classification to handle complex decision boundaries (Anbalagan et al., 2024).
- **BERT:** Leveraged the pre-trained transformer model with fine-tuning tailored separately for Tamil and Malayalam datasets, incorporating

Table 1: Data Distribution of Training and Testing Datasets

Language	Training Comments	Testing Comments
Tamil	808	100
Malayalam	800	200

advanced tokenization techniques (Ramachandruni et al., 2024).

4.2 Preprocessing Steps

To ensure data consistency and enhance model performance, the following preprocessing steps were undertaken:

- **Text Cleaning:** Removed special characters, normalized transliterated text, and addressed the challenges of code-mixing.
- **Feature Extraction:** Applied TF-IDF vectorization for LR and SVM to capture key textual patterns.
- **Tokenization:** Used BERT's subword tokenization to segment text into meaningful units for deep learning.

4.3 Training Process

The training process was designed to maximize the models' predictive capabilities:

- **Data Partitioning:** Split datasets into training and validation subsets, ensuring balanced representation of classes.
- **Model Optimization:** Conducted cross-validation and hyperparameter tuning to identify optimal configurations for each model.
- **Evaluation:** Monitored performance using accuracy and macro F1 metrics to ensure alignment with task objectives.

4.4 Deeper Analysis of SVM vs. BERT Performance in Tamil

Our experiments indicate that the SVM model outperforms BERT on Tamil data. Several linguistic and modeling factors contribute to this outcome:

- **Morphological Robustness:** Tamil's rich morphology benefits from SVM's TF-IDF n-gram representation, while BERT's subword tokenization may obscure semantics.
- **Code-Mixing and Transliteration:** SVM's bag-of-words approach is less affected by transliteration errors, whereas BERT struggles with out-of-vocabulary terms.
- **Dataset Limitations:** Tamil's limited dataset hinders BERT's fine-tuning, as its pretraining favors high-resource languages with standard orthography.

5 Experimental Results

The experimental evaluation reveals the performance of the classification models as summarized in Table 2. Among the models, SVM-Tamil achieved the highest accuracy (89.0%) and Macro F1 score (89.0%), demonstrating its robustness in classifying Tamil AI-generated and human-written reviews. LR-Tamil followed closely with an accuracy of 88.27% and an F1 score of 89.14%. in handling the intricate linguistic features of Tamil and Malayalam. BERT (Bala and Krishnamurthy, 2023) demonstrated competitive performance but faced challenges with code-mixed (Hande et al., 2021) and nuanced text.

The confusion matrix helps identify systematic misclassification trends.The confusion matrices corresponding to each model and language are presented in Figures 2 to 7, providing a detailed breakdown of predictions for human-written and AI-generated reviews.

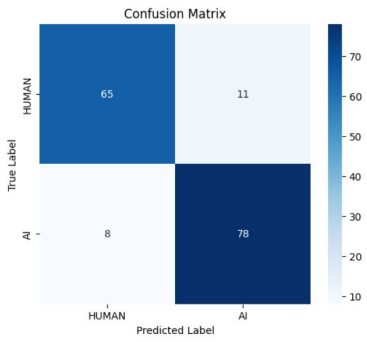

Figure 2: Confusion Matrix for LR Model (Tamil).

Table 2: Performance of models on the test dataset.

Model	Language	Accuracy	Precision	Recall	F1-Score
Logistic Regression (LR)	Tamil	88.27%	87.64%	90.70%	89.14%
Logistic Regression (LR)	Malayalam	76.88%	77.92%	75.00%	76.43%
Support Vector Machine (SVM)	Tamil	89.00%	89.00%	89.00%	89.00%
Support Vector Machine (SVM)	Malayalam	77.00%	77.00%	77.00%	77.00%
BERT	Tamil	78.00%	88.24%	62.50%	73.17%
BERT	Malayalam	79.01%	85.14%	73.26%	78.75%

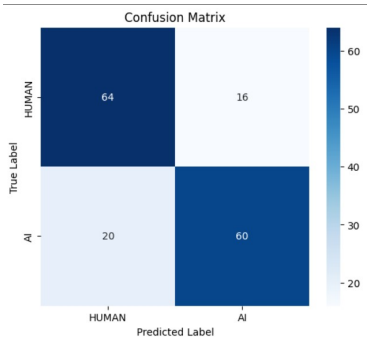

Figure 3: Confusion Matrix for LR Model (Malayalam).

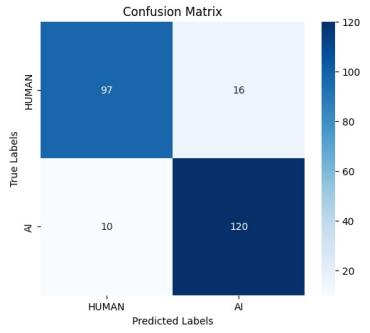

Figure 4: Confusion Matrix for SVM Model (Tamil).

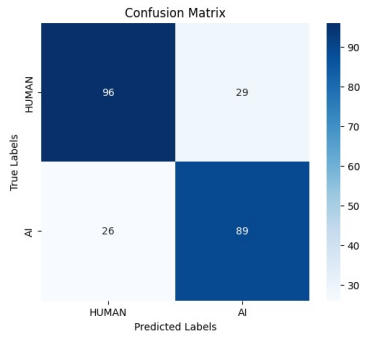

Figure 5: Confusion Matrix for SVM Model (Malayalam).

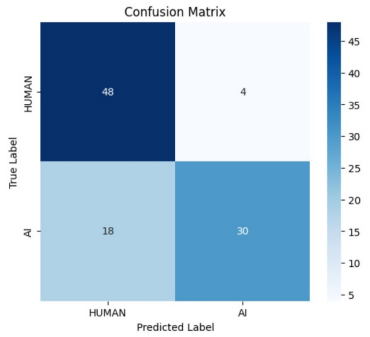

Figure 6: Confusion Matrix for BERT Model (Tamil).

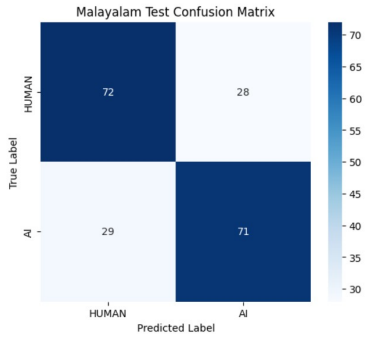

Figure 7: Confusion Matrix for BERT Model (Malayalam).

6 Conclusions

This study demonstrates the effectiveness of (Ando and Zhang, 2005) machine learning and (Bala and Krishnamurthy, 2023) deep learning models in detecting AI-generated content in Dravidian languages. BERT performed best for Malayalam, SVM for Tamil, and LR provided a strong baseline.

Future work will explore unsupervised and multilingual models to improve generalization in low-resource settings. This research advances AI-generated content detection in code-mixed languages.(Ignat et al., 2024)

For details, please visit the GitHub Repository.

7 Limitations

The model performed worse on Malayalam, achieving 79.01% accuracy with BERT, compared to Tamil, where the model reached 89.0% accuracy with SVM. Additionally, the model may misclas-

sify AI-generated text that closely mimics human writing. Another limitation is its difficulty in handling text that contains a mix of Tamil/Malayalam and English words, or text in Romanized script. Furthermore, with only approximately 800 samples per language, the model's generalization to unseen data is limited, particularly for new AI-generated or human-written reviews.

References

VP Abeera, Sachin Kumar, and KP Soman. 2023. Social media data analysis for malayalam youtube comments: Sentiment analysis and emotion detection using ml and dl models. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 43–51.

Tolulope Olalekan Abiola, Tewodros Achamaleh Bizuneh, Fatima Uroosa, Nida Hafeez, Grigori Sidorov, Olga Kolesnikova, and Olumide Ebenezer Ojo. 2025. Cic-nlp at genai detection task 1: Advancing multilingual machine-generated text detection. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 262–270.

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Akshatha Anbalagan, T Priyadharshini, A Niranjana, Shreedevi Balaji, and Durairaj Thenmozhi. 2024. Wordwizards@ dravidianlangtech 2024: Fake news detection in dravidian languages using cross-lingual sentence embeddings. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 162–166.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Abhinaba Bala and Parameswari Krishnamurthy. 2023. Abhipaw@ dravidianlangtech: Fake news detection in dravidian languages using multilingual bert. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 235–238.

D Venkata Bhargav and R Dhanalakshmi. 2024. Performance analysis of logistic regression algorithm and random forest algorithm for predicting product review analysis. In *2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, pages 1–5. IEEE.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Jose A Diaz-Garcia and Joao Paulo Carvalho. 2025. A survey of textual cyber abuse detection using cutting-edge language models and large language models. *arXiv preprint arXiv:2501.05443*.

Adeep Hande, Karthik Puranik, Konthala Yasaswini, Ruba Priyadharshini, Sajeetha Thavareesan, Anbukkarasi Sampath, Kogilavani Shanmugavadivel, Durairaj Thenmozhi, and Bharathi Raja Chakravarthi. 2021. Offensive language identification in low-resourced code-mixed dravidian languages using pseudo-labeling. *arXiv preprint arXiv:2108.12177*.

Oana Ignat, Xiaomeng Xu, and Rada Mihalcea. 2024. Maide-up: Multilingual deception detection of gpt-generated hotel reviews. *arXiv preprint arXiv:2404.12938*.

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. IEEE.

Ishika Joshi, Ishita Gupta, Adrita Dey, and Tapan Parikh. 2024. 'since lawyers are males..': Examining implicit gender bias in hindi language generation by llms. *arXiv preprint arXiv:2409.13484*.

Samsun Knight, Yakov Bart, and Minwen Yang. 2023. Generative ai and user-generated content: Evidence from online reviews. *Northeastern U. D'Amore-McKim School of Business Research Paper*, (4621982).

Balázs Kovács. 2024. The turing test of online reviews: Can we tell the difference between human-written and gpt-4-written online reviews? *Marketing Letters*, pages 1–16.

Kirti Kumari, Shirish Shekhar Jha, Zarikunte Kunal Dayanand, and Praneesh Sharma. 2023. Ml&ai_iiitranchi@ dravidianlangtech: Leveraging transfer learning for the discernment of fake news within the linguistic domain of dravidian language. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 198–206.

Olumide E Ojo, Olaronke O Adebanji, Hiram Calvo, Alexander Gelbukh, Anna Feldman, and Ofir Ben Shoham. 2024. Doctor or ai? efficient neural network for response classification in health consultations. *IEEE Access*.

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.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Sajeetha Thavareesan, Dhivya Chinnappa, Durairaj Thenmozhi, and Rahul Ponnusamy. 2021. Overview of the dravidiancodemix 2021 shared task on sentiment detection in tamil, malayalam, and kannada. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 4–6.

Indusree Ramachandruni, Sourav Mondal, Naga Venkata Mani Charan Jaladhi, Modukuri Sai Vyshnavi, Venkata Sai Sudheer Kumar Batchu, and Debnarayan Khatua. 2024. Enhancing product review authenticity detection with ensemble learning and bert model. In *2024 First International Conference on Innovations in Communications, Electrical and Computer Engineering (ICICEC)*, pages 1–6. IEEE.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Mahammad Khalid Shaik Vadla, Mahima Agumbe Suresh, and Vimal K Viswanathan. 2024. Enhancing product design through ai-driven sentiment analysis of amazon reviews using bert. *Algorithms*, 17(2):59.