AiMNLP@DravidianLangTech 2025:
Unmask It! AI-Generated Product Review Detection in Dravidian Languages

Somsubhra De and **Advait Vats**

Indian Institute of Technology Madras

Correspondence: somsubhra@outlook.in

Abstract

The rise of Generative AI has led to a surge in AI-generated reviews, often posing a serious threat to the credibility of online platforms. Reviews serve as the primary source of information about products and services. Authentic reviews play a vital role in consumer decision-making. The presence of fabricated content misleads consumers, undermines trust and facilitates potential fraud in digital marketplaces. This study focuses on detecting AI-generated product reviews in Tamil and Malayalam, two low-resource languages where research in this domain is relatively under-explored. We worked on a range of approaches - from traditional machine learning methods to advanced transformer-based models such as Indic-BERT, IndicSBERT, MuRIL, XLM-RoBERTa and Malayalam-BERT. Our findings highlight the effectiveness of leveraging the state-of-the-art transformers in accurately identifying AI-generated content, demonstrating the potential in enhancing the detection of fake reviews in low-resource language settings.

Keywords: AI-Generated review detection, classification, Dravidian Languages, NLP, Transformers, IndicSBERT, MuRIL, Malayalam-BERT

1 Introduction

In recent years, rapid advancements in artificial intelligence (AI) have significantly transformed various domains, including online content generation. Among these, the rise of AI-generated product reviews has become a major concern. These reviews, often hard to tell apart from human-written ones, threaten the trust and reliability of online platforms by influencing consumer opinions and disrupting market fairness. Since most consumers rely heavily on reviews before purchasing a product, it is essential that they differentiate between human-written and AI-generated reviews before coming to a decision. Investigations have identified apps with thousands of five-star ratings, many of which are convincingly crafted by AI. A 2023 analysis of around a million reviews revealed that 25% of top apps in popular categories on Google Play and 17% on the iOS App Store had suspicious reviews. Double Verify's Fraud Lab reported a threefold increase in apps with AI-powered fake reviews in 2024 compared to the same period in 2023 (Koetsier, 2024). In response to the growing issue, companies like Amazon have stated that it is using advanced AI to detect inauthentic product reviews (Economic-Times, 2023).

While AI-powered review detection has advanced significantly for English, research in Dravidian languages remains limited. With a growing number of online shoppers relying on local-language reviews, there is a clear need for effective detection systems. The challenge is further amplified by the prevalence of code-mixed content such as Tamil written in Roman script, English words in Tamil script, intra-sentential switching, etc. - which is common in product reviews. Variations in linguistic features including syntax, morphology, lexicon, make the process more complex. Developing robust detection systems for Dravidian languages could help address these challenges & better serve the Dravidian community. Such systems could serve as a valuable use case for integration into e-retail platforms, thereby improving transparency and trust in online marketplaces.

This study[1] contributes to the domain in the following aspects:

- We explore a range of ML, DL and SoTA transformer models to determine effective methods for detecting AI-generated reviews in the given dataset.

- We analyze & provide insights into the strengths, drawbacks of each model & perform a detailed error analysis.

---

[1]The data & codes are publicly available at https://github.com/somsubhra04/dravlangtech_ai-gen-prod-rev

2 Related Work

With LLMs like ChatGPT becoming commonplace, human and AI-generated texts are increasingly blending together in areas such as news, reviews, and social media, making it increasingly harder to distinguish between them as LLMs continue to improve. The study (Fraser et al., 2024) examines the data collection process for datasets used in AI and human-generated text detection, referencing (Su et al., 2024; TUM, 2023), pointing out that these are carefully curated and controlled rather than being organically sourced from online sources. The issue ties back to the fundamental challenge of the lack of a tool capable of definitively distinguishing between these two types of text. As a result, researchers have had to rely on pre 2020(before widespread adoption of LLMs) texts as human-labeled data and generate AI text themselves. The study further notes that AI text detection tools such as GPTZero, Originality.ai and CopyLeaks exist, but none of these provide a definitive solution at this point. Instead, the most reliable approach, recommended in the study, is to aggregate the various tools' result to obtain a reliable outcome. Given that this is the state of detection for English, the challenge is even greater for low-resource languages like Tamil and Malayalam, where even fewer datasets are available for research.

While quite a few studies have focused on NLP in Dravidian languages (Chakravarthi et al., 2023, 2024), the application to product reviews is still a relatively new area. There are few studies that specifically focus on human and AI-text detection in Tamil or Malayalam. A related study, with some similarity (Farsi et al., 2024) focused on the detection of fake news in Malayalam, where Task-1 involved the binary classification of the news into original or fake category. It explored various models, ranging from ensemble methods to deep learning and transformer-based models. Although it was the transformer models that achieved the highest performance- MuRIL-BERT, Indic-SBERT, and XLM-R recorded the highest F1 score of 0.86. In contrast, the deep learning model performed least effectively.

(Singhal and Bedi, 2024) used XLM-RoBERTa-large for a multi-class sentiment analysis of code-mixed Tamil, where it achieved an F1-score of 0.21. The study identified it as their best performing model. The model was then fine tuned for 20 epochs, with maximum sequence length of 512, using the Adam optimizer and cross entropy loss as the loss function.

Similarly, while deep learning models had the least performance in (Farsi et al., 2024), (He et al., 2017) showcased that a combination of BiLSTM-CNN gave a higher F1-score than individual DL models, when used on a dataset comprising of English tweets.

3 Task & Dataset Description

This binary classification task aims to distinguish between two categories of reviews: AI-generated (*AI*) and human-written (*HUMAN*). The detailed distribution of the datasets provided by (Premjith et al., 2025) is presented in Table 1. As observed, the label distribution is even, with no signs of class imbalance.

Language	Dataset	Classes		Total
		AI	HUMAN	
Tamil	Train	405	403	808
	Test	48	52	100
Malayalam	Train	400	400	800
	Test	100	100	200

Table 1: Dataset distribution for Tamil and Malayalam product reviews

4 Methodology

4.1 Pre-processing & Feature Extraction for DL & ML Approach

The following steps were applied to the raw text: *data cleaning* (HTML tags, punctuation, digits and extra whitespaces were removed using regular expressions), *tokenization* (text was tokenized into words for word-level analysis), and label encoding (target labels were converted into numerical labels using LabelEncoder). For *feature extraction*, **TF-IDF** (TfidfVectorizer transformed the text into numerical vectors representing the importance of words in the document, with up to 5000 features including unigrams and bigrams) and **Word2Vec** embeddings (a Word2Vec model trained on the text data generated 100-dimensional word vectors, and an average vector was computed for each document) were applied. These features were combined into a single matrix to provide a richer representation of the text data and *scaled* using StandardScaler to normalize the values, enhancing the performance of scale-sensitive models.

4.2 Model Training

4.2.1 Transformers

We applied several pre-trained transformer models which include **Indic-BERT** (Kakwani et al., 2020), **IndicSBERT** (Deode et al., 2023), **MuRIL** (Khanuja et al., 2021), **XLM-RoBERTa** (Conneau et al., 2020) and **Malayalam-BERT** (Joshi, 2023). AI4Bharat's Indic-BERT (Bidirectional Encoder Representations from Transformers) is a multilingual transformer model pre-trained on 12 major Indic languages, designed to capture language-specific nuances. The Indic sentence BERT (IndicSBERT) is a simplified variant of BERT tailored for 10 Indian languages, optimized for sentence-level understanding. Google's MuRIL (Multilingual Representations for Indian Languages)-base-cased is a transformer model trained on a large corpus of text data from 17 Indian languages, enhancing both language understanding and contextual embeddings. XLM-R is a multilingual transformer model built on the RoBERTa architecture, designed for cross-lingual understanding. It is trained on a massive amount of data across 100 languages, making it highly effective for various multilingual NLP tasks. Malayalam-BERT is a pre-trained transformer model specifically fine-tuned for Malayalam.

For each model, the *AutoTokenizer* from the Hugging Face[2] library was used to tokenize the text data automatically based on the specific model's architecture. The maximum sequence length was set to 128, and a *batch size of 16* was used. The models were fine-tuned on 80% of the train set (with *random state = 42*) for *3 epochs* with a *learning rate of* $2e^{-5}$ and weight decay of 0.01. We've used Google Colab free version T4 GPU for running the experiments. Weights & Biases (wandb) was used for experiment tracking, logging metrics and visualizing model performance during training.

4.2.2 DL Models

CNN+BiLSTM: The Convolutional Neural Network (CNN) layer captures local patterns and n-grams with 128 filters and a kernel size of 5, applying ReLU activation to introduce non-linearity - this layer helps the model learn spatial features from the input text. The following MaxPooling1D layer reduces the dimensionality, helping to retain only the most significant features. The Bidirectional Long Short-Term Memory (BiLSTM) layer

captures both forward and backward dependencies in the text, helping to understand word context. Dropout layers with a rate of 0.5 are applied after the LSTM and Dense layers for regularization. Finally, a *GlobalAveragePooling1D* layer reduces the BiLSTM output to a fixed-size vector, which is then passed through a Dense layer with 64 units and ReLU activation. The output layer is a Dense layer with softmax activation to produce class probabilities for multi-class classification. The model was trained on 80% of the train set using the *Adam optimizer* (*learning rate:* $1e^{-3}$) and sparse categorical cross-entropy loss over *15 epochs* for both the Tamil & Malayalam datasets respectively with a *batch size of 32*.

4.2.3 Traditional Approaches

We trained Support Vector Machine (SVM) using Grid Search with 5-fold cross-validation to find the optimal combination of the hyper-parameters (kernel types - *'linear', 'rbf', 'poly', 'sigmoid'*, regularization parameter C values: *[0.1, 1, 10, 100]* & kernel coefficient gamma - *'scale', 'auto'*). Also, Random Forest with 100 estimators was trained on the feature set. XGBoost classifier, a gradient boosting algorithm, was trained with 100 estimators and learning rate of 0.1. We then combined both RF and XGBoost using a VotingClassifier for a soft voting.

5 Results

5.1 Quantitative Analysis

The *macro-avg. F1-score*[3] is utilized as the primary metric to assess the overall effectiveness of the system. Table 2 presents a detailed comparison of the performance across all models and approaches evaluated in this study.

For Task-1 (Tamil), **IndicSBERT outperforms** all models with the highest F1-score (96%). IndicSBERT builds on multilingual BERT by fine-tuning it for cross-lingual sentence representation learning. This simple yet effective approach without explicit cross-lingual training enhances its ability to capture linguistic properties across languages.

Indic-BERT, MuRIL and XLM-RoBERTa demonstrate strong results but slightly lower than IndicSBERT. The Random Forest and XGBoost ensemble approach shows a relatively promising result however struggles with the Malayalam dataset,

---

[2]https://huggingface.co

[3]https://scikit-learn.org/1.5/modules/generated/sklearn.metrics.f1_score.html

which might hint at challenges with the complexity of Malayalam. CNN+BiLSTM performs decently but lags behind transformer models. Malayalam-BERT outperforms in the Malayalam task with an impressive 92% F1-score. This improved version of BERT was fine-tuned by (Joshi, 2023) on publicly available monolingual Malayalam datasets, as existing multilingual models did not perform as well on downstream tasks. IndicSBERT is again the top performer along with MuRIL. Interestingly, the XLM-RoBERTa-base model achieved a perfect precision & recall (1.0) for the human & AI classes respectively, resulting in zero false positives for human class and zero false negatives for AI-generated text. This means the model accurately identified all AI-generated texts. Whenever a text was classified as human-written, the prediction was always correct. Such performance is typically seen in imbalanced datasets where the model tends to favor the dominant class. However, this was not the case here, as both the training & test data samples had only two instances extra, from either class. The model made errors in six specific cases (shown in fig 3). In each case, the model misclassified human-written texts as AI-generated. This suggests that while the model could learn distinct patterns in AI-generated data, human written texts with their more diverse styles, might have been harder to categorize. The MuRIL model also demonstrated a similar trend during validation, where it produced comparable results.

Coming to the DL models and traditional methods, the results shown in table 6 indicate that the SVM model did well on both datasets, with macro-F1 scores of 0.85 and 0.77 in validation. However, its performance dropped on the test set, indicating some overfitting. The CNN+BiLSTM model showed a contrasting trend, with a relatively lower validation F1 but a significant improvement on the Tamil test set. However, the model struggled with generalization in Malayalam. The ensemble classifier experienced a performance drop in Malayalam on the test set, while maintaining strong results in Tamil.

5.2 Qualitative Analysis

We analyzed the characteristics of AI-generated and human-written product reviews on the train & test sets (Tables 3, 4), focusing on their linguistic differences, common patterns & sources of misclassification. A clear difference between the two categories is their length and complexity.

In Malayalam, AI-generated reviews have a lower average word count compared to human-written reviews. A similar trend is observed in sentence length where AI-generated reviews tend to have shorter and more direct sentences compared to human reviews. However, in Tamil, this pattern is reversed - AI-generated reviews are significantly longer. This suggests that AI-generated content in Tamil may be overly descriptive compared to human reviews, which are often brief and to the point.

Interestingly, AI-generated Malayalam reviews have higher lexical diversity than human-written ones. This suggests that AI-generated reviews may use a broader vocabulary or introduce uncommon words that are less typical in natural user reviews. The opposite is observed in Tamil, where human reviews show higher lexical diversity compared to AI-generated reviews. We analyzed the false positives and false negatives for both tasks. A key question arises: *For common misclassifications, which models are performing better & predicting correctly?* Figures 8, 9 show all reviews that were misclassified by more than one transformer model. ✓ indicates that the label has been correctly predicted by the model, while ✗ denotes incorrect prediction. For eg., in the 6th Tamil review, only XLM successfully identifies the AI-generated content, while all other models fail in this case.

We conducted a brief survey where 19 individuals proficient in Tamil or Malayalam reviewed misclassified samples. They were asked to categorize each sample as AI-generated or human-made based solely on their judgment, without access to ground truth labels or using any translation tools. Respondents who answered the survey most accurately observed the following: For Tamil, AI-generated text often uses uncommon words in multiple sentences, with some original Tamil words that have transitioned to colloquial usage. For Malayalam, they identified grammatical errors, unusual word choices, incorrect word placement, tense errors, unrelated words and lack of sentence continuity as indicators of AI-generated text.

6 Conclusion

The performance of various transformer and DL models was examined for classifying product reviews into Human written and AI generated categories for low resource languages like Tamil and Malayalam. While the DL models performed some-

Model	Task-1 (Tamil)				Task-2 (Malayalam)			
	P	**R**	**F1**	**Acc.**	**P**	**R**	**F1**	**Acc.**
Indic-BERT	0.93	0.93	0.93	0.93	0.86	0.85	0.85	0.85
IndicSBERT	0.96	0.96	**0.96**	0.96	0.92	0.92	0.91	0.92
MuRIL	0.94	0.94	0.94	0.94	0.9	0.9	0.9	0.9
XLM-RoBERTa	0.94	0.94	0.94	0.94	0.87	0.87	0.87	0.87
Malayalam-BERT	-				0.93	0.93	**0.92**	0.93
CNN+BiLSTM	0.89	0.89	0.89	0.89	0.69	0.6	0.55	0.6
SVM	0.77	0.77	0.77	0.77	0.65	0.65	0.65	0.65
Ensemble (RF+XGBoost)	0.9	0.9	0.9	0.9	0.6	0.59	0.59	0.59

Table 2: Performance of various models on the test-set (macro-averaged Precision, Recall, F1 and Accuracy scores from the *best run* for each approach have been mentioned)

what promisingly when multiple models were combined, their performance still did not match the performance of transformer models, especially for Malayalam, highlighting their strong capability & efficacy in handling complex linguistic features in the Dravidian space. Among the best performing models were-Indic-BERT, IndicSBERT, MuRIL and XLM-RoBERTa. In general, all the models achieved a lower F1-score on Malayalam samples than Tamil, with Malayalam-BERT producing the best results for Malayalam classification. IndicS-BERT performed best on the Tamil samples and closely followed Malayalam-BERT for Malayalam samples, making it the *most efficient model when evaluated across both languages.*

6.1 Future Work

Future work will focus on employing LLMs (few-shot, CoT prompting, exploring RAG), trying ensemble methods with transformer-based models & expanding to other low-resource languages for cross-lingual transfer learning. Additionally, we plan to conduct experiments on larger and more diverse datasets as the current study was limited in scope. This will help reproduce and assess the real-world applicability of our models, ensuring their effectiveness at scale. Also, one critical concern is the risk of misclassifying human-written text as AI-generated, leading to false positives. This can have significant consequences such as unwarranted censorship & questioning of genuine user feedback. Moreover, the ethical implications of AI-generated text detection need to be considered, particularly regarding privacy and bias. An important future direction will be ensuring that the detection systems developed are both accurate and fair, minimizing the chances of misclassification.

7 Limitations

The transformer models, pre-trained on corpora created for different tasks, may limit their performance on review detection. The lack of sufficient data in low-resource languages hampers effective fine-tuning for this specific task. Additionally, the dataset used was not code-mixed. Compute limitations restricted the ability to fine-tune transformers efficiently. Also, we could have analyzed the misclassified examples to check for any possible bias in the transformers, which might have given useful insights. Specifically, it would have been important to examine whether the model shows biases towards certain groups of reviews, such as favoring specific dialects or writing patterns/styles. However, due to our lack of proficiency in Tamil and Malayalam, we were unable to carry out an in-depth analysis.

Acknowledgments

References

Bharathi R. Chakravarthi, Ruba Priyadharshini, Anand Kumar M, Sajeetha Thavareesan, and Elizabeth Sherly, editors. 2023. *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*. INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Anand Kumar Madasamy, Sajeetha Thavareesan,

Elizabeth Sherly, Rajeswari Nadarajan, and Manikandan Ravikiran, editors. 2024. *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics, St. Julian's, Malta.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Samruddhi Deode, Janhavi Gadre, Aditi Kajale, Ananya Joshi, and Raviraj Joshi. 2023. L3cube-indicsbert: A simple approach for learning cross-lingual sentence representations using multilingual bert. *Preprint*, arXiv:2304.11434.

Economic-Times. 2023. Using advanced ai to spot and remove fake customer reviews: Amazon. (Accessed: 22 January 2025).

Salman Farsi, Asrarul Eusha, Ariful Islam, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. CUET_Binary_Hackers@DravidianLangTech EACL2024: Fake news detection in Malayalam language leveraging fine-tuned MuRIL BERT. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 173–179, St. Julian's, Malta. Association for Computational Linguistics.

Kathleen C. Fraser, Hillary Dawkins, and Svetlana Kiritchenko. 2024. Detecting ai-generated text: Factors influencing detectability with current methods. *Preprint*, arXiv:2406.15583.

Yuanye He, Liang-Chih Yu, K. Robert Lai, and Weiyi Liu. 2017. YZU-NLP at EmoInt-2017: Determining emotion intensity using a bi-directional LSTM-CNN model. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 238–242, Copenhagen, Denmark. Association for Computational Linguistics.

Raviraj Joshi. 2023. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *Preprint*, arXiv:2211.11418.

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.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages. *Preprint*, arXiv:2103.10730.

J. Koetsier. 2024. Fake ai-generated reviews flooding app stores. (Accessed: 22 January 2025).

B Premjith, Nandhini K, Bharathi Raja Chakravarthi, Thenmozhi Durairaj, Balasubramanian Palani, Sajeetha Thavareesan, and Prasanna Kumar Kumaresan. 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.

Kriti Singhal and Jatin Bedi. 2024. Transformers@DravidianLangTech-EACL2024: Sentiment analysis of code-mixed Tamil using RoBERTa. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 151–155, St. Julian's, Malta. Association for Computational Linguistics.

Jinyan Su, Claire Cardie, and Preslav Nakov. 2024. Adapting fake news detection to the era of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1473–1490, Mexico City, Mexico. Association for Computational Linguistics.

TUM. 2023. Idmgsp-galactica-train-cg: A fine-tuned galactica model to detect machine-generated scientific papers. Accessed: 2025-01-29.

A Appendix

Language	Metric	AI	Human
Tamil	Average Word Count	7.904	5.700
	Average Sentence Length	1.000	1.025
	Lexical Diversity	0.992	0.987
Malayalam	Average Word Count	12.155	16.815
	Average Sentence Length	1.043	1.458
	Lexical Diversity	0.983	0.939

Table 3: Analysis of AI-generated and human-written reviews for Tamil and Malayalam train-sets

Language	Metric	AI	Human
Tamil	Average Word Count	23.146	4.115
	Average Sentence Length	2.333	1.019
	Lexical Diversity	0.858	0.983
Malayalam	Average Word Count	12.05	22.57
	Average Sentence Length	1.00	1.73
	Lexical Diversity	0.996	0.921

Table 4: Analysis of AI-generated and human-written reviews for Tamil and Malayalam test-sets

	REVIEW		GROUND TRUTH
	TASK 1	**TASK 2**	
	பெரிய அளவுக்கு கரையாது இதைப் பயன்படுத்தும்போது நிறைய சோப்பைப் பயன்படுத்த வேண்டிய நிலை ஏற்படுகிறது. *(It does not dissolve in large quantities and when used, it becomes necessary to use a lot of soap.)*	ബിരിയാണി, പപ്പടം, അച്ചാർ - മറ്റെവിടെയും കിട്ടാത്ത ഒരു രുചി അനുഭവം. *(Biryani, papad, pickles - a taste experience like no other.)*	**AI**
	ஸ்கிரீன் ரொம்ப பெருசா இருக்கும் ஆனா ரொம்ப வெயிட்டான லேப்டாப் *(The screen is very large but the laptop is very heavy.)*	ബ്രൊ ഞാൻ സ്കിൻ കെയർ ഇതുവരെ ചെയ്തിട്ടില്ല എന്റെ മുഖത്തിന് ഒരു കുഴപ്പവും ഇല്ല ഞാൻ ഇതു ഫോളോ ചെയ്തിട്ട് പ്രശ്നം ആയി *(Bro, I haven't done skin care yet, my face is fine, I followed this and got a problem.)*	**HUMAN**

Figure 1: Sample Tamil and Malayalam Texts from the Training Set with English Translations* for Context
(Note* The English translations may not fully capture the nuances, sentiment and cultural context inherent in the original Tamil and Malayalam texts. As a result, the English version might not reflect the true tone or intention.)
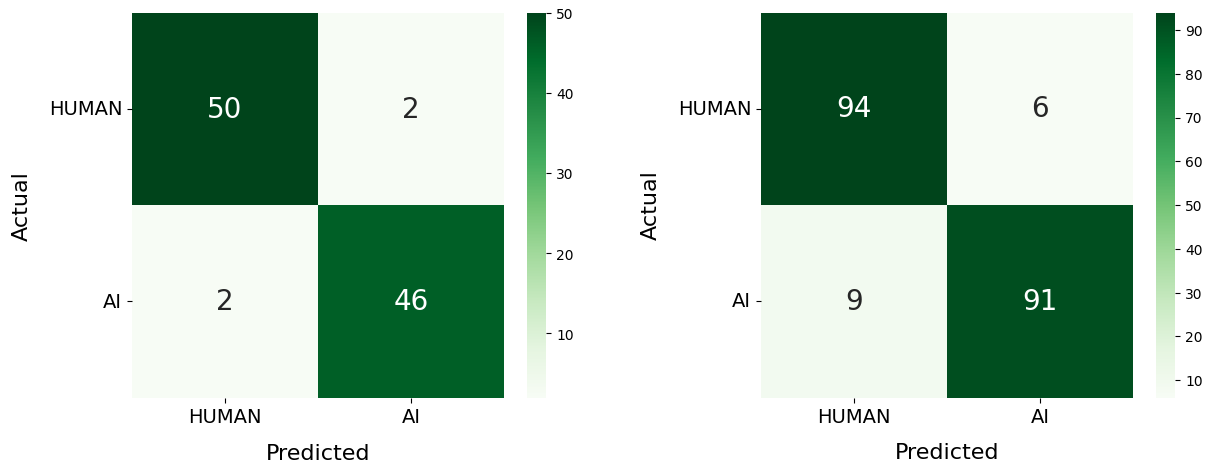

Figure 2: Confusion Matrix for the best runs on Tamil & Malayalam test sets (using IndicSBERT & Malayalam-BERT respectively)
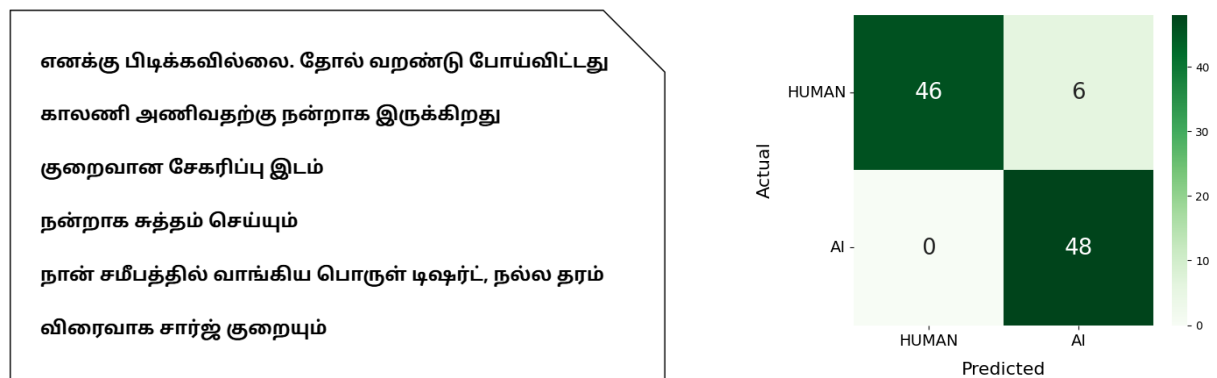

Figure 3: Misclassified Tamil texts from the test set alongside the confusion matrix: XLM-R incorrectly predicted these 6 human-written reviews as AI-generated.

172

Model	Category	Task-1 (Tamil)				Task-2 (Malayalam)			
		P	R	F1	Acc.	P	R	F1	Acc.
Indic-BERT	HUMAN	0.93	0.93	0.93	-	0.94	0.93	0.93	-
	AI	0.94	0.94	0.94	-	0.94	0.93	0.93	-
	macro avg.	0.94	0.94	0.94	0.94	0.93	0.93	0.93	0.93
IndicSBERT	HUMAN	0.97	0.99	0.98	-	0.95	0.95	0.95	-
	AI	0.99	0.98	0.98	-	0.95	0.95	0.95	-
	macro avg.	0.98	0.98	0.98	0.98	0.95	0.95	0.95	0.95
MuRIL	HUMAN	1	0.97	0.99	-	0.97	0.91	0.94	-
	AI	0.98	1	0.99	-	0.92	0.97	0.95	-
	macro avg.	0.99	0.99	**0.99**	0.99	0.95	0.94	0.94	0.94
XLM-RoBERTa	HUMAN	1	0.86	0.92	-	1	0.84	0.91	-
	AI	0.89	1	0.94	-	0.86	1	0.92	-
	macro avg.	0.94	0.93	0.93	0.93	0.93	0.92	0.92	0.92
Malayalam-BERT	HUMAN			-		0.99	0.95	0.97	-
	AI			-		0.95	0.99	0.97	-
	macro avg.			-		0.97	0.97	**0.97**	0.97

Table 5: Performance *(Label-wise scores from the classification report)* of transformer-based models on the **validation set**

Model	Category	Task-1 (Tamil)				Task-2 (Malayalam)			
		P	R	F1	Acc.	P	R	F1	Acc.
CNN+BiLSTM	HUMAN	0.67	0.55	0.6	-	0.91	0.50	0.65	-
	AI	0.66	0.76	0.7	-	0.66	0.95	0.78	-
	macro avg.	0.66	0.65	0.65	0.66	0.78	0.72	0.71	0.72
SVM	HUMAN	0.78	0.95	0.86	-	0.81	0.7	0.75	-
	AI	0.94	0.77	0.85	-	0.74	0.84	0.78	-
	macro avg.	0.86	0.86	**0.85**	0.85	0.77	0.77	**0.77**	0.77
Ensemble (RF+XGBoost)	HUMAN	0.78	0.8	0.79	-	0.72	0.65	0.68	-
	AI	0.82	0.8	0.81	-	0.68	0.75	0.71	-
	macro avg.	0.8	0.8	0.8	0.8	0.7	0.7	0.7	0.7

Table 6: Performance *(Label-wise scores from the classification report)* of DL & ML approaches on the **validation set**

Figure 4: Most common words in AI-generated *(left)* & human-written *(right)* reviews on the **Tamil train** set

173

Figure 5: Most common words in AI-generated *(left)* & human-written *(right)* reviews on the **Tamil test** set

Figure 6: Most common words in AI-generated *(left)* & human-written *(right)* reviews on the **Malayalam train** set

Figure 7: Most common words in AI-generated *(left)* & human-written *(right)* reviews on the **Malayalam test** set

Text	True Label	MuRIL	IndicBERT	IndicSBERT	XLM
எனக்கு பிடிக்கவில்லை. தோல் வறண்டு போய்விட்டது.	HUMAN	✗	✗	✗	✗
காலணி அணிவதற்கு நன்றாக இருக்கிறது.	HUMAN	✗	✗	✗	✗
குறைவான சேகரிப்பு இடம்.	HUMAN	✗	✗	✓	✗
நன்றாக சுத்தம் செய்யும் .	HUMAN	✗	✓	✓	✗
விரைவாக சார்ஜ் குறையும்.	HUMAN	✗	✗	✓	✗
நான் அண்மையில் வாங்கிய ஒரு வாட்டர் பாட்டில் மிகவும் அருமையாக இருக்கின்றது. அது மிகவும் ஸ்டைலிஷ், எளிதில் பயன்படுத்த கூடியது,	AI	✗	✗	✗	✓
நான் சமீபத்தில் வாங்கிய பொருள் டிஷர்ட், நல்ல தரம்.	HUMAN	✓	✗	✓	✗

Figure 8: Common misclassified reviews in Tamil

Text	True Label	MuRIL	IndicBERT	IndicSBERT	XLM	MALBERT
മലയാറ്റൂർ രാമകൃഷ്ണന്റെ യക്ഷി എന്ന നോവൽ മലയാളത്തിലെ മികവുറ്റ ഒരു സൈക്കോളജിക്കൽ ത്രില്ലർ ആണ്.	HUMAN	✗	✗	✗	✗	✗
ജോപ്പന്റെ "ദൈവത്തിന്റെ ചാരന്മാർ" വായിച്ചിട്ടുള്ളവർക്ക് തീർച്ചയായിട്ടും ഈ അഛരന്റെ രൂപം മനസ്സിൽ പതിഞ്ഞിട്ടുണ്ടാവും.	HUMAN	✗	✗	✗	✗	✗
ഈ കാർ ടാക്സി പോലുള്ള വാണിജ്യ ആവശ്യങ്ങൾക്ക് ഉപയോഗിക്കാൻ കഴിയില്ല...	HUMAN	✗	✓	✗	✗	✗
ഫ്രണ്ടിൽ ലോവർ സിൽവർ ട്രിം കാണാൻ വൃത്തികേട് ; ബാക്കിയെല്ലാം കൊള്ളാം; ഇന്റീരിയർ സൂപ്പർ .	HUMAN	✗	✓	✓	✗	✗
മീറ്റർ കണ്ടോൾ ചെയ്യാൻ ഹാൻഡിലിൽ ജോയ്സ്റ്റിക്; ചിലപ്രമുഖ കാർ കമ്പനികൾ പോലും മീറ്ററിൽ തന്നെ ഞെക്കാൻ കുറ്റി നൽകുമ്പോൾ ആണ് ഇതെന്ന് ഓർക്കണം.	HUMAN	✗	✗	✓	✗	✗
നല്ല നേട്ടം തരുന്ന ഇൻവെസ്റ്റ്മെന്റ് അറിയാത്ത ആളുകൾക്ക് കേൾക്കുമ്പോൾ നല്ലത് പോലെ തോന്നിക്കുമെങ്കിലും, ഇത് അത്ര നല്ല പ്ലാൻ അല്ല. നിങ്ങളെ പോലെ ഉള്ള ഏജന്റ്മാർ പൊതുവെ വിൽക്കാൻ നോക്കാറ് ഇതുപോലെ ഉള്ള സ്കീമുകളാണ്, കാരണം ഇതിൽ ഏജന്റ് കമ്മിഷൻ വളരെ ഉയർന്നതാണ്.	HUMAN	✗	✓	✗	✗	✗
യഥാർത്ഥത്തിൽ എഴുതപ്പെട്ട പല പുസ്തകങ്ങൾ സമൂഹത്തോട് പ്രതികരിക്കാനായി ജീവിക്കുന്നു	AI	✗	✗	✓	✓	✗
40 ലക്ഷം പറഞ്ഞിട്ട് 80 ലക്ഷം വരെ ആഗ്രഹിച്ചാൽ അതാണല്ലോ പിള്ളേരുടെ കാഴ്ചപ്പാട്	AI	✗	✗	✗	✗	✗
ഇക്കാലത്ത് ഒക്കെ ടാറ്റയുടെ ഡിസൈൻ കണ്ട് കണ്ണിനു ഊട്ടുപോലും ഇടുന്നുണ്ട്	AI	✗	✗	✗	✗	✗
ഇന്റെര്യർ കല്ലും കൊടിയും പോല ഡിസൈൻ ചെയ്തതാണെന്ന് തോന്നും, എന്തിനാണു ഇങ്ങനെയൊക്കെ?	AI	✗	✗	✓	✓	✓
മാരുതി എള്ളുപ്പം സെയിൽ അടിച്ചാൽ പോലും ഇത്രയും ഓപ്ഷൻസ് തരണമെന്നില്ല	AI	✗	✗	✗	✗	✗
സെൽറ്റോസ് കാറിന് എന്തൊരു ഗ്രൗണ്ട് ക്ലിയറൻസ്; ഇങ്ങനെ താഴെ ചെറുതായെങ്കിലും തടസ്സം ഉണ്ടാക്കിയാൽ എങ്ങിനെ മണ്ണിടിഞ്ഞു പോവില്ല?	AI	✗	✗	✓	✓	✓
5 സ്റ്റാർ സേഫ്റ്റി പറഞ്ഞിട്ട് വേറെ ഒന്ന് തലയ്ക്കലുമില്ല... പാസ്സന്റജേഴ്സ് പൊറുതിയാലെ ഇനി തീരും.	AI	✗	✗	✗	✗	✗
ഇന്ത്യയിലെ റോഡുകളിലേക്കുള്ള കാർ എങ്ങനെ ഇങ്ങനെ കണ്ട് ഡിസൈൻ ചെയ്യുന്നു?	AI	✗	✗	✓	✓	✓
എനിക്ക് 2015 മോഡൽ ഡ്യൂക്ക് 390ഉം 2023 മോഡൽ അധ്വഞ്ചർ 390ഉം ഉണ്ട്... ബൈക്ക് കിടിലം ആണെങ്കിലും ഹെഡ്ലൈറ്റ് നോക്കി ചുമ്മാ പണത്തെ ഫീൽ തന്നെ.	AI	✗	✗	✗	✗	✗
ഇൻഫിനിറ്റി എക്സ്എം ഒന്ന് വെട്ടിപ്പോയി... ട്രയംഫ് എപ്പോഴും 100% പവർ ടൂ വെറ്റ് ചെയ്ത് ഇടും.	AI	✗	✗	✗	✗	✗
റോയൽ എൻഫീൽഡ് സീരിയസ് ആയി 800 സിസി ബൈക്ക് ഇറക്കിയാൽ ഇന്ത്യയിൽ ബുള്ളറ്റ് മാരുടെ ആഘോഷം ഇരട്ടിയാകും	AI	✗	✗	✗	✗	✗
നമുക്ക് കൂടി ലഭിക്കുന്ന ഇഎംഐഎക്ക് ഒരുപാട് മോഡലുകൾ എടുക്കാമെങ്കിലും എങ്ങനെ ഉപേക്ഷിക്കും പഴയ ഹീറോയുടെ വിശ്വാസ്യത?	AI	✗	✗	✓	✓	✓
പലർക്കും തോന്നും ക്രെഡിറ്റ് കാർഡ് പ്രശ്നങ്ങളുണ്ടാക്കുമെന്ന്, പക്ഷേ കൃത്യമായി പ്ലാൻ ചെയ്താൽ ഇത് നമുക്ക് സാമ്പത്തിക മാനേജ്മെന്റിനായി ഒരു നല്ല സംരക്ഷണ മാർഗമാണു; ഞാൻ പല ആവശ്യങ്ങൾക്കും ഇതിനൊപ്പം കാശുള്ളതായി ഇരിക്കാനുള്ള ഒരു ഫിനാന്ഷ്യൽ ഡിസിപ്ലിൻ കരസ്ഥമാക്കി.	AI	✗	✗	✓	✓	✓
ഇൻഷുറൻസ് ഏജന്റുമാർ വാഗ്ദാനം ചെയ്യുന്ന വളരെ മികച്ച റിട്ടേൺ സ്കീമുകൾ കേട്ടാൽ നല്ലതാണെന്ന് തോന്നും, പക്ഷേ തീർക്കുന്ന സമയത്ത് ഒന്നും തന്നെ കൈയിലെത്തില്ല; ഏജന്റിന് മാത്രമേ ഇതിൽ ഗുണമുള്ളൂ	AI	✗	✗	✓	✗	✗
ഉറുബി ന്റെ സുന്ദരി കളും സുന്ദരന്മാരും മികച്ച നോവൽ ആയി ആദ്യ പരിഗണന യിൽ വരേണ്ട തായിരുന്നു. അത് കഴിഞ്ഞേ മറ്റെന്തും ഉള്ളൂ.	HUMAN	✓	✗	✓	✗	✓
പുതിയ ഡെസൈൻ വന്നിട്ട് ഹെഡ്റസ്റ്റ് ഇപ്പഴും നാല് മാത്രം കൊടുക്കുന്ന മാരുതി തന്നെയാണ്.	AI	✓	✗	✗	✓	✓
റോയൽ എൻഫീൽഡ് 650ന്റെ സൗണ്ട് ഒരു തലമറയ്ക്കും ശബ്ദം.. എന്നാൽ 6 ലക്ഷം പണത്തിന് ക്യാശ് വെച്ചാൽ എക്സ്ട്രാ ഫീച്ചറുകൾ വേണം.	AI	✓	✗	✗	✓	✓
എഴുത്തും വായനയും ലോപിച്ച് പോകുന്നീക്കാലത്ത് കേശവദേവിന്റെ നോവൽ മുതൽ വർത്തമാനകാല സൃഷ്ടികൾ വരെ ഉൾപ്പെടുന്ന ലിസ്റ്റ്! ഇനിയും ധാരാളം ബാക്കി!	HUMAN	✓	✓	✗	✗	✓

Figure 9: Common misclassified reviews in Malayalam