# RATHAN@DravidianLangTech 2025: Annaparavai - Separate the Authentic Human Reviews from AI-generated one

**Jubeerathan Thevakumar**
Dept. of Computer Sci. and Eng
University of Moratuwa
Colombo, Sri Lanka
jubeerathan.20@cse.mrt.ac.lk

**Luheerathan Thevakumar**
Jaffna, Sri Lanka
the.luheerathan@gmail.com

## Abstract

Detecting AI-generated reviews is crucial for maintaining the authenticity of online feedback in low-resource languages like Tamil and Malayalam. We propose a transfer learning-based approach using embeddings from XLM-RoBERTa, IndicBERT, mT5, and Sentence-BERT, validated with five-fold cross-validation via XGBoost. These embeddings are used to train deep neural networks (DNNs), refined through a weighted ensemble model. Our method achieves 90% f1-score for Malayalam and 73% for Tamil, demonstrating the effectiveness of transfer learning and ensembling for review detection. The source code is publicly available to support further research and improve online review systems in multilingual settings.

## 1 Introduction

As artificial intelligence technologies become increasingly sophisticated, the proliferation of AI-generated reviews presents a growing threat to the integrity of online consumer feedback systems. Recent studies have revealed that a significant portion of reviews in sectors such as home services, legal, and medical fields are likely fraudulent, with many confirmed as AI-generated (Karaş, 2024; Thilagavathi et al., 2024). These fake reviews undermine consumer trust, create unfair competition, and pose significant challenges for e-commerce platforms and consumers alike. The rapid production of convincing fake reviews threatens the foundational trust mechanism of online marketplaces, necessitating robust detection systems and enhanced consumer protection measures to maintain the integrity of online review ecosystems.

This research focuses on detecting AI-generated product reviews in Tamil and Malayalam, two low-resource languages spoken in South India. The increasing presence of fraudulent online reviews in these languages underscores the urgent need for effective detection methods. However, the scarcity of linguistic resources and tools for these low-resource languages presents significant challenges. To mitigate these limitations, we utilize two datasets introduced by (Premjith et al., 2025), which comprises both AI-generated and human-authored product reviews in Tamil and Malayalam.

We employed a transfer learning-based approach for feature extraction, utilizing embeddings from four different models. These embeddings are evaluated through cross-validation using XGBoost to assess their discriminative capacity. Following this, we train four independent deep neural network (DNN) models on the extracted embeddings. Finally, we construct an ensemble model that aggregates predictions from the individual models, aiming to improve classification performance through weighted averaging. Our implementation is publicly available in the GitHub[1] repository.

The findings from this study have important implications for strengthening content moderation systems in e-commerce platforms, ultimately fostering greater transparency and trust in online review ecosystems.

## 2 Related Work

The task of detecting AI-generated product reviews is a subset of the broader AI-generated text detection challenge. While most research in this area has focused on widely spoken languages, there is a notable lack of studies addressing AI-generated content in Tamil and Malayalam.

[1] https://github.com/Jubeerathan/Annaparavai



*Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 449–453

(Ippolito et al., 2019) employed a set of BERT-based classifiers (Devlin et al., 2019) with three popular random decoding strategies—top-k, nucleus, and temperature sampling—on text samples generated by GPT-2 (Radford et al., 2019). (Fagni et al., 2021) introduced a set of sequence-based classifiers, including LSTM, GRU, and CNN, for detecting AI-generated social media texts.

RoBERTa, a pretrained, non-generative language model (Liu, 2019), was integrated into classifiers to detect text generated by GPT-2 (Solaiman et al., 2019). Despite having a distinct architecture and tokenizer compared to GPT-2, the RoBERTa-based classifier was able to detect text generated by the GPT-2 model with an accuracy of approximately 95%.

Stylometric features, which are quantitative characteristics of a person's writing style, can be used alongside pre-trained language models to enhance detection capabilities. These features highlight the stylistic differences between human and AI authors, aiding in the detection of AI-generated text. Incorporating stylometric aspects such as phraseology, punctuation, and linguistic diversity into pre-trained language model-based classifiers has shown improved performance in detecting AI-generated tweets (Kumarage et al., 2023). Ensemble learning techniques, combined with stylometric features, Linguistic Word Inquiry, GPT-2 word embeddings, and Author's Multilevel Ngram Profiles (AMNP) features, are utilized alongside transfer learning (Mikros et al., 2023) to identify the AI-generated text.

Similar to stylometric features, other notable efforts have focused on leveraging various text characteristics to enhance detection capabilities. SeqXGPT, for example, uses sentence-wise log probability metrics from white-box LLMs to identify AI-generated text at the sentence level (Wang et al., 2023). GPT-who revisits the Uniform Information Density (UID) hypothesis, proposing that AI-generated text may lack the evenness in information distribution typical of human language, and introduces UID features to measure the smoothness of token distribution (Venkatraman et al., 2023). Additionally, another approach improves detection accuracy by combining the factual structure of text with a RoBERTa-based classifier (Zhong et al., 2020). These methods utilize structural and sequential features to enhance the detection of AI-generated content.

Most of these studies collectively underscore the critical role of transformer-based architectures in addressing the challenges of detecting AI-generated content, especially in the English language. By refining language-specific models and exploring multimodal techniques, these research efforts have created a solid groundwork for future progress in the field of AI-generated content detection.

## 3 Dataset

We used two data sets for this investigation: the Tamil and Malayalam datasets from (Premjith et al., 2025). The Tamil dataset consists of 808 samples in the training set, and 100 samples in the given test set. The Malayalam dataset contains 800 samples in the training set, and 200 samples in the given test set. Both training datasets are annotated with labels Human and AI.

Figures 1 and 2 illustrate the length distribution of the training datasets and testing datasets of each language.

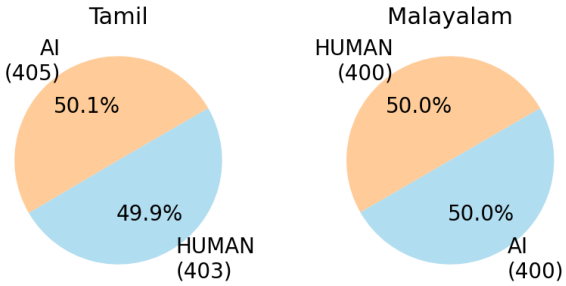


Figure 1: Distribution of labels in Tamil and Malayalam languages in train dataset.

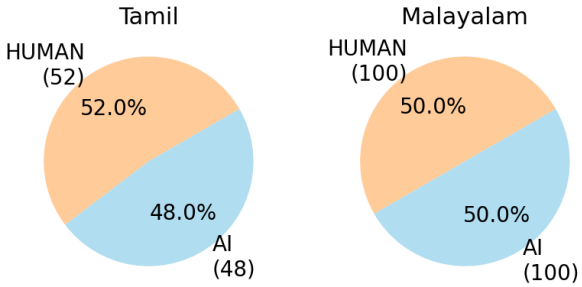


Figure 2: Distribution of labels in Tamil and Malayalam languages in given test dataset.

## 4 Methodology

### 4.1 Data Preprocessing

We did not require any data preprocessing steps for our dataset, as it consists of short texts with a max-

imum of 2 sentences and 3-4 words per sentence. Each of these sentences is clean and consistent, adhering to a standardized format. This high level of data quality means that there are no spelling errors, grammatical mistakes, or irrelevant content that would necessitate additional cleaning or normalization.

Furthermore, the language models we used to generate the embeddings, such as indic-bert (Kakwani et al., 2020), are designed to handle a variety of text inputs and perform certain preprocessing tasks internally. These models are capable of tokenizing the text, managing special characters and punctuation, and adjusting the length of text inputs through padding and truncation. This built-in preprocessing capability of the language models ensures that minor inconsistencies or noise in the data are effectively managed, allowing us to generate high-quality embeddings without the need for extensive data cleaning steps.

In summary, the combination of a clean and consistent dataset with the robust preprocessing capabilities of the language models we employed allowed us to bypass additional data preprocessing steps, streamlining our workflow and ensuring efficient and accurate text embedding generation.

### 4.2 Model Training

In our research, we aimed to detect AI-generated product reviews in Tamil and Malayalam by leveraging the strengths of monolingual models. We designed two monolingual models, one for Tamil and one for Malayalam.

First, we generated embeddings for each text entry in our dataset using a variety of language models, including XLM-RoBERTa (Conneau, 2019), Indic-BERT (Doddapaneni et al., 2023), and mT5 (Xue, 2020), which were trained on various languages, specifically on Tamil and Malayalam. Additionally, we employed Sentence-BERT (Reimers, 2019), which has been effectively used for AI-generated or AI-paraphrased text detection (Schaaff et al., 2024). These embeddings captured the semantic and syntactic properties of the text, providing a rich representation for further analysis.

To evaluate the effectiveness of each model's embeddings, we performed five-fold cross-validation using XGBoost. This ensured that our feature representations were robust across different subsets of the dataset.

Next, we split the dataset into three parts: 70% for training, 21% for testing, and 9% for validation. The training set was used to train the individual Deep Neural Network (DNN) models, while the testing set was used to evaluate their performance. We trained four separate DNN models independently using embeddings from each language model. To enhance overall performance, we employed a weighted average ensembling technique, leveraging the complementary strengths of different models.

Our evaluation metric was the F1-score, which provided a balanced measure of precision and recall, ensuring a more reliable assessment of classification performance compared to accuracy. By training the DNN with these features, we developed a streamlined and efficient model suitable for low-resource environments while maintaining strong classification performance in detecting AI-generated product reviews in Tamil and Malayalam.

Figure 3 illustrates the architecture of the model and figure 4 shows the detailed methodology of the work.

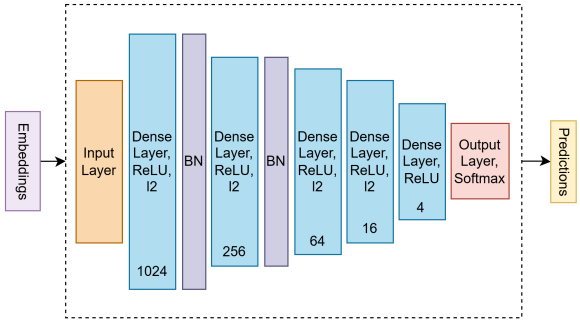


Figure 3: Proposed DNN architecture.

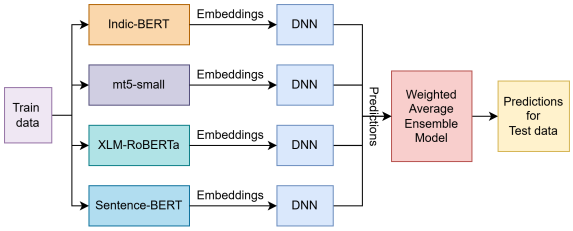


Figure 4: Proposed methodology.

## 5 Results and Discussion

Our initial XGBoost model achieved promising results on a 5-fold cross-validation set. Table 1 and Table 2 shows the mean and the standard deviation for each individual models. F1-score produced by

| | Mean | Std.Dev |
|---|---|---|
| Sentence-BERT | 0.962 | 0.014 |
| XLM-RoBERTa | 0.948 | 0.013 |
| Indic-BERT | 0.965 | 0.010 |
| mT5 | 0.945 | 0.013 |

Table 1: Cross validation set mean and standard deviation for Tamil dataset.

| | Mean | Std.Dev |
|---|---|---|
| Sentence-BERT | 0.934 | 0.015 |
| XLM-RoBERTa | 0.909 | 0.010 |
| Indic-BERT | 0.927 | 0.011 |
| mT5 | 0.930 | 0.010 |

Table 2: Cross validation set mean and standard deviation for Malayalam dataset.

the proposed DNN and ensemble models are shown in Table 3.

For the given test set, the ensemble model achieved 0.73 for Tamil and 0.90 for Malayalam. Respective confusion matrices are shown in the Figure 5 and Figure 6.

Despite achieving promising results with XGB on the cross-validation set and the proposed models on splitted test set, we observed a performance drop on the given test set for Tamil. This discrepancy may be attributed to differences in the distributions of the training and test sets, potentially generated by different LLM models. Future efforts will focus on refining the ensemble DNN model to ensure uniformity across varying distributions.

## 6 Conclusion

In this study, we explored AI-generated product review detection in Tamil and Malayalam using monolingual models with transfer learning and ensembling. Our approach achieved 90% accuracy for Malayalam and 73% for Tamil, demonstrating the effectiveness of transfer learning in low-resource Dravidian languages.

| Models | Tamil | Malayalam |
|---|---|---|
| Sentence-BERT DNN | 0.959 | 0.905 |
| XLM-RoBERTa DNN | 0.971 | 0.964 |
| Indic-BERT DNN | 0.971 | 0.935 |
| mT5 DNN | 0.953 | 0.964 |
| Ensemble model | 0.982 | 0.940 |

Table 3: Predictions of proposed models for 21% split of train dataset in Tamil and Malayalam.

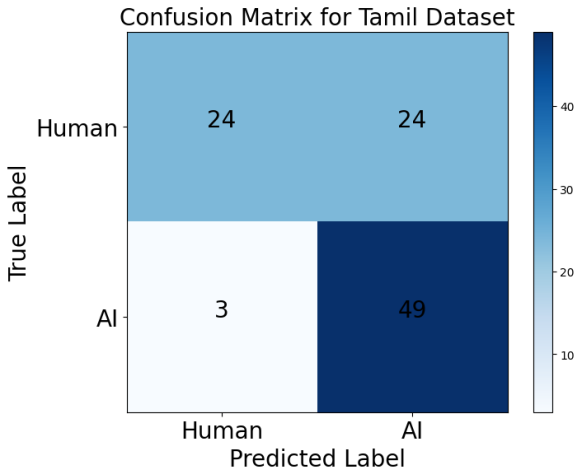


Figure 5: Confusion matrix for Tamil.

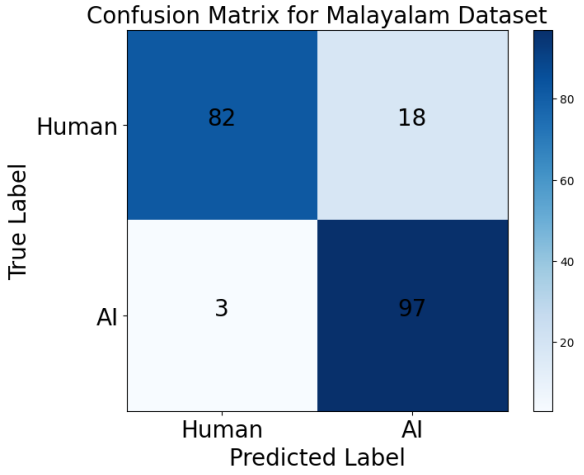


Figure 6: Confusion matrix for Malayalam.

To support research in this field, we have made our source code publicly available, enabling replication and further development. This contribution fosters innovation and collective efforts to enhance the reliability of AI-generated content detection, promoting the integrity of online reviews.

## 7 Limitations

The AI-generated reviews in the dataset may exhibit biases inherited from the language models used to generate them. These biases could affect the performance and fairness of the detection model, leading to variations in effectiveness. Additionally, the dataset is limited, which may further constrain the model's ability to learn. Addressing both dataset limitations and inherent biases remains a crucial area for future research.

## References


A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. Tweepfake: About detecting deepfake tweets. *Plos one*, 16(5):e0251415.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2019. Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint arXiv:1911.00650*.

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 EMNLP*.

Zeynep Karaş. 2024. Effects of ai-generated misinformation and disinformation on the economy. *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, 12(4):2349–2360.

Tharindu Kumarage, Joshua Garland, Amrita Bhattacharjee, Kirill Trapeznikov, Scott Ruston, and Huan Liu. 2023. Stylometric detection of ai-generated text in twitter timelines. *arXiv preprint arXiv:2303.03697*.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

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*.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Kristina Schaaff, Tim Schlippe, and Lorenz Mindner. 2024. Classification of human-and ai-generated texts for different languages and domains. *International Journal of Speech Technology*, 27(4):935–956.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

K Thilagavathi, K Thankamani, P Shunmugapriya, and D Prema. 2024. Navigating fake reviews in online marketing: Innovative strategies for authenticity and trust in the digital age. *The Scientific Temper*, 15(03):2854–2858.

Saranya Venkatraman, Adaku Uchendu, and Dongwon Lee. 2023. Gpt-who: An information density-based machine-generated text detector. *arXiv preprint arXiv:2310.06202*.

Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang, and Xipeng Qiu. 2023. Seqxgpt: Sentence-level ai-generated text detection. *arXiv preprint arXiv:2310.08903*.

L Xue. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Wanjun Zhong, Duyu Tang, Zenan Xu, Ruize Wang, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Neural deepfake detection with factual structure of text. *arXiv preprint arXiv:2010.07475*.