CIC-NLP@DravidianLangTech 2025: Detecting AI-generated Product Reviews in Dravidian Languages

Tewodros Achamaleh[1], Abiola T. O.[1], Lemlem Eyob[1], Mikiyas Mebiratu[2], Grigori Sidorov[1]
[1]Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico City, Mexico
[2]Wolkite University, Department of Information Technology, Wolkite, Ethiopia

Abstract

AI-generated text now matches human writing so well that telling them apart is very difficult. Our CIC-NLP team submits results for the DravidianLangTech@NAACL 2025 shared task to reveal AI-generated product reviews in Dravidian languages. We performed a binary classification task with XLM-RoBERTa-Base using the DravidianLangTech@NAACL 2025 datasets offered by the event organizers. By training the model effectively, our experiments distinguished between human and AI-generated reviews with scores of 0.96 for Tamil and 0.88 for Malayalam in the evaluation test set. This paper presents detailed information about preprocessing, model architecture, hyperparameter fine-tuning settings, the experimental process, and the results. The source code is available on GitHub.[1]

1 Introduction

The fast growth of Large Language Models (LLMs) now changes how natural language processing works across many uses (Yigezu and Tesfaye, 2023; Kolesnikova and Ivanov, 2023; Adebanji and Okoro, 2024; García-Vázquez and Rodriguez, 2023; Laureano and Calvo, 2024; Aguilar-Canto and Ramirez, 2023; Ojo and Bello, 2024; Brown and Leike, 2023; Abiola et al., 2025b,a). Computer algorithms make Machine-generated text through AI while showing human writing ability with limited human involvement. MGT has revolutionized production through automated content creation, but labs now must excel at recognizing MGT from HWT texts, especially in situations demanding proof like product evaluation.

Human authors create text that harnesses personal experiences to comprehend cultures and emotions, which allows them to present detailed feelings that fit perfectly into their context. According to (Zhang et al., 2024), MGT shows language precision at its surface level but fails to achieve the contextual synergy present in HWT. Underrepresented Dravidian languages show distinct characteristics that make their interpretation different from other languages (Conneau et al., 2020; Ruder et al., 2023).

Detecting MGT is essential for stopping online lies and resolving ethical issues with AI-generated content (Ansarullah, 2024; Floridi and Cowls, 2023). Language models built at scale need training data that holds stereotypes to produce outputs that follow established verbalization patterns(Gallegos et al., 2024; Brennan and Greenstadt, 2023). These computing system prejudices create analytical opportunities to tell actual human-written text from machine-generated text through detailed language marker inspection. Our team joins the DravidianLangTech@NAACL 2025 Shared Task to create AI-generated product review detection systems for Tamil and Malayalam. Our work involved differentiating AI-generated and human-written reviews across Tamil and Malayalam using an exceptional data resource that includes multiple language forms from human writers and computer systems. Our research used XLM-RoBERTa-Base, a transformer model for multilingual text understanding (Liu and Ott, 2023), as the basis for our experiment. Our research confirms how the model understands varied language styles and shows why different data sets need separate treatment in AI content detection technology.

Our methodology achieved macro average F1 scores of 00.96 for Tamil and 0.88 for Malayalam on the evaluation test set. This paper's main contribution is to provide insights into preprocessing, model architecture, hyperparameter tuning, and evaluation. It correspondingly contributes to the growing AI content detection research in low-resource languages. Our work brings to the forefront the potential of fine-tuned multilingual

[1]https://github.com/teddymas95/AI-generated-Product-Reviews

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 580–585
May 3, 2025

models for NLP in underrepresented languages by addressing the complexity of multilingual AI-generated texts.

2 Related Work

From here to the research, the focus was on finding clear indicators of AI-generated content through pattern detection or verbalization inconsistencies (Maimone and Jolley, 2023; Aydin and Kara, 2023; Clark et al., 2023). Nevertheless, with the development of generative models regarding text generation quality and contextual coherence (Smith et al., 2023; Brown et al., 2024), machine-generated text increasingly became more complicated to differentiate. As these advancements were made, traditional rule-based systems became not adequate, pushing us into the field of deep learning approaches, in particular using transformer-based models (Kierner et al., 2023; Chen and Wang, 2024; Jurafsky and Martin, 2023). Natural language processing (NLP) has come a long way, but transformer models have greatly improved it. Several studies have shown them to be very strong at NLP tasks such as sentiment analysis, text classification, and data summarization (Soto et al., 2024; Hoang, 2024; Zhang et al., 2024; Ruder et al., 2023).

(Gupta and Verma, 2023) XLM-RoBERTa was consistently widely preferred for Multilingual tasks, especially in low-resource languages like Tamil and Malayalam, due to its strong cross-lingual performance (Conneau et al., 2020). In particular, (Li et al., 2024) studied the issue of how to build robust AI detection systems for varied text types and multiple language models. The paper emphasized the need to deal with text variability in real-world text and showed how named entities and structural details may help identify differences between AI-generated and human-written text, but with slight differences as AI systems improve (Brennan et al., 2023). (Fernández-Hernández et al., 2023; Eyob et al., 2024) Performed a series of experiments with multilingual BERT for the AuTexTification shared task at IberLEF 2023 concerning distinguishing AI-generated texts (García-Vázquez and Rodriguez, 2023). Their findings demonstrated that fine-tuned transformer models could outperform traditional machine-learning techniques without including metadata features like readability and sentiment.

Also, (Kumar et al., 2024) looked at how well hybrid transformer-based architectures deal with linguistic diversity, specifically in classifying texts from several domains (Joshi et al., 2024). These studies have been restricted to high-resource languages, and it remains challenging to detect machine-generated text (MGT) in low-resource languages like Tamil and Malayalam (Chatterjee et al., 2023; Kumar et al., 2023; Achamaleh et al., 2024). As linguistically diverse and morphologically complex as Hindi is, and scarce are large annotated datasets, tailored methods are called. This work proceeds prior work using XLM-Roberta in its multilingual capabilities, developing the detection of AI-generated content in Dravidian languages and outperforming state of the art.

3 AI vs. Human Text Detection

3.1 Dataset Analysis

The organizers provided datasets for training and testing data through Google Drive(Premjith et al., 2025). Each dataset consists of the following columns: ID, DATA, and LABEL. The Label column contains two values: The datasets classify text as HUMAN when humans compose it and AI when AI systems produce it. Our primary objective is to differentiate AI-generated text from human-written content. The Tamil dataset includes 808 records of AI-generated (405 texts) and human-written material (403 texts). The Malayalam dataset provides 800 texts made by both AI generators and humans, with 400 texts in each group. The team made this dataset to represent normal content variations in real-world data, supporting high-quality model testing and training. During this task, the datasets were split into training, validation, and testing sets, enabling the fine-tuning of our XLM-RoBERTa-Base model. The balanced class distribution in the datasets contributed to achieving reliable and unbiased model performance across Tamil and Malayalam.

3.2 XLM-RoBERTa-Base

We used the Transformers Library from Hugging Face and fine-tuned the XLM-RoBERTa Base, a multilingual transformer model for binary Tamil and Malayalam text classification. To prepare and format the dataset and satisfy the model's input needs, we respected specific tokenization and inherent linguistic caveats about these languages. Our team then processed the dataset by passing it through the XLM-RoBERTa tokenizer to prepare for training and testing. Our method included

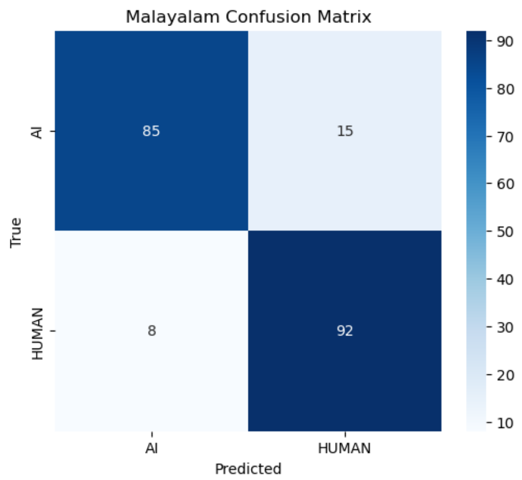

Figure 1: Malayalam Confusion Matrix

adding dropout to prevent model overtraining and saving checkpoints during training to pick the version that did best according to validation metrics. We adopted a cosine-annealing learning rate schedule to stabilize training and improve the final model performance. This paper provides the methodology of adapting to use XLM-RoBERTa-Base effectively for this task. Section 4 explains how we preprocessed the dataset, encoded the text, and were ready for further analysis.

4 System Setup and Experiments

4.1 System Setup

We trained the XLM-RoBERTa-Base model as a multilingual transformer architecture tuned to identify pairs of binary classes in Tamil and Malayalam language datasets. The datasets were preliminary processed by Hugging Face AutoTokenizer, which turned text entries into model-ready tokenized content. Hugging Face datasets library divided our information into 90% training data and 10% test data sets. To recognize text types, the XLM-RoBERTa model required the addition of a classification segment that generated human or AI predictions. With a learning rate of 3e-5, we trained our model across five epochs using batches of 8 and regularized dropout layers to avoid overfitting. Our system selected the optimal model results by evaluating performance on early stop conditions and checkpointed models. We assessed model performance using F1 scores, precision, recall, and accuracy during test dataset predictions.

4.2 Experiments

For the binary classification of human- and AI-generated text in Tamil and Malayalam datasets, we fine-tuned the XLM-RoBERTa-Base model. This multilingual transformer design took in both dataset characteristics well. The model achieved better results by selecting specific values for important training settings such as batch size, learning rate, and training steps. Our model used GPU computing during five training epochs and divided gradient updates into two steps to fit memory. We took advantage of the mixed precision training to make training run faster and to prevent overfitting by early stopping based on the validation of the F1 score. Moreover, a cosine-annealing learning rate scheduler and warmup steps stabilized the training, with the learning rate starting low and increasing gradually at the beginning and becoming lower after some time.

DataCollatorWithPadding was used to dynamically pad input sequences for each batch to the maximum sequence length for computational efficiency. This reduced the number of extra operations on padding tokens, which made the model more attentive to meaningful text content. F1-score and loss metrics were used closely to indicate the training and validation performance. The results include training and validation plots, which show that the model achieved competitive performance with macro F1 scores. The results indicate the robustness of the fine-tuning approach and the selection of good hyperparameters. After each epoch, we evaluated the model's performance on the development dataset, tracking its progress and ensuring the training and validation metrics were aligned. This helped identify potential issues such as overfitting or underfitting early in the process.

5 Results

Our evaluation tests the performance of our fine-tuned XLM-RoBERTa-Base model across Tamil and Malayalam datasets for a binary classification setup. We evaluated model performance by running text predictions on development data and measured accuracy plus micro and macro F1 scores. Our Tamil model achieved 0.96 accuracy as measured by macro F1 scores to differentiate content created by AI from human producers. The dataset balance and rich vocabulary influenced Tamil text, giving rise to this excellent model performance. Even with uneven class distribution in Malayalam data,

Language	Model	Precision	Recall	F1-Score	Accuracy
Malayalam	xlm-roberta-base	0.9739	0.9739	**0.9739**	0.975
	distilbert-base-uncased	0.9194	0.9271	0.9226	0.925
	bert-base-multilingual-cased	0.9479	0.9479	0.9479	0.950
Tamil	xlm-roberta-base	0.9509	0.9286	**0.9358**	0.9383
	distilbert-base-uncased	0.9423	0.9143	0.9225	0.9259
	bert-base-multilingual-cased	0.9509	0.9286	0.9258	0.9283

Table 1: Model Comparison for Malayalam and Tamil on the Development Dataset.

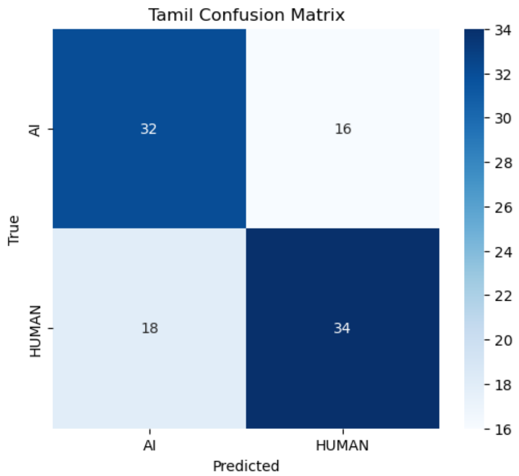

Figure 2: Tamil Confusion Matrix

the model delivered an impressive 0.88 Macro F1 score, demonstrating its ability to work across multiple languages.The model shows a strong ability to differentiate AI and human-generated text in Tamil and Malayalam with reliable detection accuracy.

6 Discussion

The detection of AI-generated reviews using XLM-RoBERTa in Tamil and Malayalam was highly effective. Training data balance for the Tamil model strengthened its generalization capability. The class imbalance in Malayalam did not affect its ability to maintain high generalization performance. Achieving multilingual NLP success depends on synchronous data quality management and proper class balancing capabilities, making transformer models ideal for low-resource language processing.

XLM-RoBERTa outperformed DistilBERT and BERT-multilingual in precision and F1-score. DistilBERT was efficient but misclassified many authentic reviews, while BERT-multilingual had uneven results, especially with Malayalam. XLM-RoBERTa showed reliable performance, though all models struggled with unclear cases. Future improvements could include domain-specific training and features like readability scores and syntactic analysis. Table 1 compares the models for both languages.

6.1 Error Analysis

The minority classes in the imbalanced Malayalam dataset showed most of the misinterpreted classification labels. AI-generated reviews in Tamil easily fooled human scrutiny because they were presented as if written by human writers. The number of wrong classifications in Malayalam increased because the language uses intricate sentence formats and blends two different written systems. Research results indicate that it is necessary to improve algorithmic models by introducing language elements that exceed simple token recognition processes. Figures 1 and 2 show the confusion matrix.

Conclusion

The research examined the power of transformer-based models to find AI-generated product reviews across the two Dravidian languages, Tamil and Malayalam. The XLM-RoBERTa model achieved better results, particularly in Tamil, since its balanced dataset helped it improve generalization abilities. The Malayalam model demonstrated robustness even though its performance was affected by the class imbalance problem. The analysis of misclassification errors during testing showed that AI mistaken instances mainly occurred when minority classes contained text similar to actual human writing. XLM-RoBERTa performed best among all three models during comparison tests because it delivered maximum precision and F1-score measurements for both language codes. All produced models encountered difficulties when classifying ambiguous instances, suggesting enhanced improvements through linguistic features must be implemented. Properly selecting high-quality multilingual datasets plays a critical role in successful NLP tasks. The future development of AI-generated text detection in low-resource languages requires research on syntactic feature integration, semantic feature integration, domain-specific fine-tuning, and metadata-based improvement methods.

Limitations

This study faced several challenges because class imbalance negatively influenced the performance of the Malayalam model. The model encountered difficulties with generalization because the dataset had a limited capacity to handle different writing styles. The models failed to function correctly while processing ambiguous cases with AI text similar to human writing. Future improvements must concentrate on growing more enormous datasets with balanced distribution and developing advanced linguistic elements to improve detection precision levels.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

Tolulope O. Abiola, Tewodros A. Bizuneh, Oluwatobi J. Abiola, Temitope O. Oladepo, Olumide E. Ojo, Adebanji O. O., Grigori Sidorov, and Olga Kolesnikova. 2025a. Cic-nlp at genai detection task 1: Leveraging distilbert for detecting machine-generated text in english. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, Abu Dhabi, UAE. Association for Computational Linguistics.

Tolulope O. Abiola, Tewodros A. Bizuneh, Fatima Uroosa, Nida Hafeez, Grigori Sidorov, Olga Kolesnikova, and Olumide E. Ojo. 2025b. Cic-nlp at genai detection task 1: Advancing multilingual machine-generated text detection. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, Abu Dhabi, UAE. Association for Computational Linguistics.

Tewodros Achamaleh, Lemlem Kawo, Ildar Batyrshini, and Grigori Sidorov. 2024. Tewodros@ dravidianlangtech 2024: Hate speech recognition in telugu codemixed text. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 96–100.

Tunde Adebanji and Chima Okoro. 2024. Ethical implications of ai in generating human-like text. *AI Society*, 39:567–580.

Diego Aguilar-Canto and Sofia Ramirez. 2023. Challenges in detecting machine-generated text in under-resourced languages. *Language Resources and Evaluation*, 57:145–161.

M. et al. Ansarullah. 2024. Inceptor regulates insulin homeostasis. *Nature Metabolism*. Referencing Gallegos et al., 2024.

Mehmet Aydin and Elif Kara. 2023. Advancements in detecting ai-generated texts: Challenges and methodologies. *AI and Society*.

M. Brennan, S. Afroz, and R. Greenstadt. 2023. Forensic linguistics for ai text detection.

M. Brennan and R. Greenstadt. 2023. Linguistic markers for ai text detection.

T. Brown and J. Leike. 2023. The sociotechnical impact of large language models.

T. Brown, B. Mann, and N. Ryder. 2024. Gpt-4: Scaling generative text quality.

Rahul Chatterjee et al. 2023. Challenges in nlp for low-resource languages: A focus on tamil and malayalam. *Low-Resource NLP Journal*.

Yuxin Chen and Jie Wang. 2024. Transformer-based architectures in modern nlp. *NLP Research Journal*.

E. Clark, A. Gupta, and K. Lee. 2023. Early detection methods for ai-generated text.

A. Conneau, K. Khandelwal, and N. Goyal. 2020. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Lemlem Eyob, Tewodros Achamaleh, Muhammad Tayyab, Grigori Sidorov, and Ildar Batyrshin. 2024. Stress recognition in code-mixed social media texts using machine learning. *International Journal of Combinatorial Optimization Problems and Informatics*, 15(1):32.

Ana Fernández-Hernández et al. 2023. Autextification shared task at iberlef 2023: Experiments with multilingual bert for ai text detection. In *Proceedings of IberLEF 2023*. Springer.

L. Floridi and J. Cowls. 2023. Ethical governance of generative ai.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79.

Elena García-Vázquez and Pedro Rodriguez. 2023. Bias in machine-generated text: A case study on multilingual models. *International Journal of Artificial Intelligence*, 32:45–62.

Rishi Gupta and Ananya Verma. 2023. Cross-lingual applications of xlm-roberta in low-resource nlp tasks. *Journal of Computational Linguistics*.

T. Hoang. 2024. Transformer models in multilingual nlp. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 456–465.

P. Joshi, S. Santy, and A. Budhiraja. 2024. Challenges in low-resource language nlp.

D. Jurafsky and J. Martin. 2023. From rule-based systems to transformers: A survey of nlp paradigms.

Tobias Kierner et al. 2023. Transformer models in natural language processing: A survey of advancements and applications. *Computational Linguistics Today*.

Anna Kolesnikova and Sergey Ivanov. 2023. Exploring multilingual text representations with transformer models. *Transactions of the ACL*, 11:212–230.

Arjun Kumar et al. 2024. Hybrid transformer-based architectures for multilingual text classification. *Journal of Artificial Intelligence and Language Technologies*.

R. Kumar, S. Murugesan, and B. Rajendran. 2023. Dravidian language processing: Trends and gaps.

Miguel Laureano and Ana Calvo. 2024. Language models and their role in sentiment analysis. *Journal of Sentiment Analysis*, 18:321–337.

Wei Li et al. 2024. Challenges in ai-detection systems for multilingual text classification. *Multilingual AI Journal*.

Y. Liu and M. Ott. 2023. Xlm-roberta for multilingual understanding.

John Maimone and Sarah Jolley. 2023. Identifying ai-generated content: Pattern detection and verbalization inconsistencies. *Journal of Artificial Intelligence Research*.

Adewale Ojo and Fatima Bello. 2024. Cross-lingual learning: Advancements in text classification. *Journal of Cross-lingual NLP*, 25:89–105.

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.

S. Ruder, M. Peters, and S. Swayamdipta. 2023. Domain adaptation in transformer models.

Hannah Smith et al. 2023. Overcoming the challenges of detecting advanced ai-generated text. *Journal of Machine Learning Applications*.

R. Soto et al. 2024. Applications of xlm-roberta in multilingual text analysis. *Transactions on Computational Linguistics*, 12:234–245.

Alemayehu Yigezu and Meron Tesfaye. 2023. The advancements in cross-lingual nlp applications. *Journal of Computational Linguistics*, 49:102–119.

Wei Zhang et al. 2024. Evaluation of transformer models in multilingual sentiment analysis tasks. *Sentiment Analytics Quarterly*.