Overview of the Shared Task on Detecting AI Generated Product Reviews in Dravidian Languages: DravidianLangTech@NAACL 2025

Premjith B[1], K Nandhini[2], Bharathi Raja Chakravarthi[3], Durairaj Thenmozhi[4], Balasubramanian Palani[5], Sajeetha Thavareesan[6], Prasanna Kumar Kumaresan[7],
[1]Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India,
[2]School of Mathematics and Computer Sciences, Central University of Tamil Nadu, India,
[3]School of Computer Science, University of Galway, Ireland,
[4]Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu, India,
[5]Indian Institute of Information Technology Kottayam, Kerala, India,
[6]Department of Computing, Eastern University, Sri Lanka,
[7]Data Science Institute, University of Galway, Ireland

Abstract

The detection of AI-generated product reviews is critical due to the increased use of large language models (LLMs) and their capability to generate convincing sentences. The AI-generated reviews can affect the consumers and businesses as they influence the trust and decision-making. This paper presents the overview of the shared task on Detecting AI-generated product reviews in Dravidian Languages" organized as part of DravidianLangTech@NAACL 2025. This task involves two subtasks—one in Malayalam and another in Tamil, both of which are binary classifications where a review is to be classified as human-generated or AI-generated. The dataset was curated by collecting comments from YouTube videos. Various machine learning and deep learning-based models ranging from SVM to transformer-based architectures were employed by the participants.

1 Introduction

Customers consider using the reviews posted on social media and e-commerce platforms to purchase products, read books, and watch entertainment shows such as movies and dramas. The reviews directly influence the economic outcomes of the businesses and create perceptions about them (Luo et al., 2023; Jabeur et al., 2023; Tiwari et al., 2024). The emergence of large language models (LLMs) and their capacity to produce human-like text raises significant concerns about the authenticity of reviews, given their potential to generate both positive and negative feedback. Users can use LLM-based tools to fabricate reviews, mislead customers, and ultimately impact businesses. Therefore, it has become crucial to distinguish between human and AI-generated reviews, a task that is challenging due to the high quality of the text generated by these AI tools. Below are some examples of AI-generated reviews [1]:

- **Example 1**: Love this! Well made, sturdy, and very comfortable. I love it!Very pretty
- **Example 2**: It's very hard to get a pair of these pants in the stores. They are a bit too small and too tight.

It is crucial to devise methodologies to detect AI-generated product reviews to ensure the authenticity of the online customer feedback. Numerous studies have been reported regarding the detection of AI-generated product reviews, ranging from traditional classifiers to advanced neural network architectures. (Wani et al., 2024) achieved an accuracy of 98.46% using a hybrid BiLSTM-Word2Vec model, while (Lee et al., 2022) and (Venugopala et al., 2024) used machine learning classifiers such as random forest and support vector machine (SVM). Though neural models (BiLSTM, GRU) often outperform traditional methods in NLP tasks, the absence of standardized benchmarks complicates direct comparisons. High accuracy claims, such as (Wani et al., 2024)'s 98.46%, risk overfitting concerns unless validated on diverse, real-world datasets. (Fayaz et al., 2020) address robustness via an ensemble model, yet the efficacy of majority voting hinges on the diversity and strength of base classifiers, a nuance not thoroughly explored. A critical issue lies in dataset heterogeneity. Studies use disparate sources (Amazon reviews, restaurant reviews) of varying sizes. Most datasets are English-only, limiting insights into

[1]https://osf.io/tyue9/

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 65–74

multilingual detection. (Gambetti and Han, 2023) introduce a GPT-based model focused on linguistic complexity, a promising feature absent in other works, but their analysis lacks cross-validation against existing methods. The reliance on static embeddings like Word2Vec restricted the performance, as newer embeddings better capture semantic nuances. While these studies demonstrate technical proficiency, real-world applicability remains underexplored. (Gambetti and Han, 2023)'s linguistic complexity analysis offers a novel direction but requires integration with behavioral or metadata features for holistic detection. While current methods show promise, their fragmentation across datasets and techniques underscores the need for cohesive frameworks. In addition, gold-standard corpora and models are unavailable for low-resource languages, such as Dravidian languages. Bridging these gaps will be essential to developing robust, adaptable solutions for AI-generated review detection.

The shared task on "Detecting AI-generated product reviews in Dravidian languages: DravidianLangTech@NAACL 2025" offers an avenue for the researchers to differentiate the human and AI-generated reviews in Malayalam and Tamil languages. This is the first instance of conducting a shared task specifically for these two languages. The shared task has two tasks—one in Malayalam and another in Tamil. This shared task introduces a novel corpus in the Malayalam and Tamil languages, which was curated by collecting comments received for YouTube review videos. We considered reviews written in Malayalam and Tamil scripts while excluding those written in Latin letters to maintain the consistency in scripts.

2 Task Description

There are two subtasks in this shared task:

- Task 1: Detecting AI-generated reviews in Malayalam
- Task 2: Detecting AI-generated reviews in Tamil

In both tasks, the objective is to classify a given review into human and AI categories.

3 Dataset Description

The dataset is prepared in Malayalam and Tamil languages. We created the dataset by maintaining an equal number of data points in the human and AI classes in both languages. This artificially created balance prevents the machine learning models from biasing toward any specific class. However, in real-world scenarios, AI-generated reviews are generally less frequent than human-written ones. Imbalanced data collected from real-world scenarios reflects the practical challenges pertaining to this task.

We prepared the dataset by considering reviews collected from various YouTube channels. Different channels were considered for preparing the training and testing data to ensure separate training and testing data distribution. We didn't consider the reviews from e-commerce platforms, and therefore we didn't conduct a cross-domain generalization test.

AI-generated reviews were found to be linguistically less complex than human-written ones. However, distinguishing between human and AI-generated reviews became difficult when AI tools were fine-tuned to mimic human writing styles effectively. Since AI-generated text can be paraphrased or structured to resemble authentic reviews, models had to rely on subtle textual features, making classification more challenging. A major challenge faced during the data collection phase was to maintain an equal number of human and AI-generated reviews. We iterated the AI review generation process multiple times to achieve class balance. We tried to reduce the bias toward any class or product by including reviews related to different product categories collected from different YouTube channels. However, we haven't addressed the problems pertaining to the gender and racial biases.

As generative AI models continue to evolve, their ability to mimic human writing will improve, which makes the differentiation harder. This dataset is the first in Malayalam and Tamil related to this task and provides a benchmark for current models. By expanding the dataset and collecting reviews from e-commerce platforms, we can enhance its generalizability and reduce its various biases.

3.1 Malayalam data

We gathered reviews from YouTube channels by categorizing various product types. To keep the train and test distribution different, we considered different sets of categories for training and testing data, and these data were collected from different YouTube channels. Table 1 shows the list of cate-

Data	Product Categories
Train	Facewash, Referral products, Dress, Makeup products, Meesho products, Furniture, Mobile phone, Movie, TV, Car, Bike, Apple Vision Pro, Laptop, Electronic gadgets, Airpod, BSNL, Internet connection, Food
Test	Food, Books, Car, Bike, Movie, Credit card, Insurance

Table 1: Categories of products used for creating the corpus

Language	Data	Class	Count	Total
Malayalam	Train	Human	400	800
		AI	400	
	Test	Human	100	200
		AI	100	
Tamil	Train	Human	403	808
		AI	405	
	Test	Human	52	100
		AI	48	

Table 2: Distribution of the data in Train and Test datasets in Malayalam and Tamil languages

gories used in training and testing data.

The flow of the dataset creation process is illustrated in Figure 1. Here, we considered two classes: human and AI. The human class includes reviews written by humans, while the AI class includes reviews generated using AI tools. Initially, the comments were collected from the YouTube videos discussing the products mentioned in Table 1. While collecting the data, we made sure that the selected comments contained only Malayalam characters. We removed all other comments during the preprocessing step. Additionally, we eliminated all emojis from the reviews. We prepared the human class data by using these reviews. We used ChatGPT to generate AI-based reviews using the approach put forward by (Xylogiannopoulos et al., 2024). During this process, we provided ChatGPT with human-written reviews and instructed it to generate 20 similar reviews in Malayalam. From the AI-generated reviews, we removed reviews containing less than 5 words and created the AI-generated review corpus. If the number of AI-generated reviews is less, we repeat the instruction until both human and AI classes have an equal number of samples. We ensured that the number of data samples for both the human and AI classes was equal.

3.2 Tamil data

The product feedback template is created using Google Forms with 20 products such as soap, shampoo, hair oil, footwear, wristwatches, mobiles, cosmetics, clothing, handbags, bikes, laptops, and so on. The individuals are advised to submit their feedback experience, both positive and negative, in Tamil without using any AI tools. We share the same template with another set of individuals and advise them to utilize various AI tools such as ChatGPT, Julius, Gemini, among others. We check user reviews for duplicates and tag it with appropriate labels such as "human written" or "AI generated". Figure 2 shows the process of Tamil dataset creation.

Table 2 explains the distribution of train and test data used in Malayalam and Tamil tasks.

4 Methodologies used in the Submissions

A total of 130 teams registered for this shared task. However, only 33 teams submitted the results in Malayalam, and 37 teams submitted them in Tamil. Some of the teams submitted multiple runs. We evaluated the submissions using the macro F1 score, and then prepared the rank list based on the results. Tables 3 and 4 show the rank lists for the Malayalam and Tamil tasks, respectively.

The descriptions of the systems used by the participating teams are given below.

4.1 Nitiz

The team implemented the multilingual AI-generated text detection model using the IndicBERT transformer, which employs a multimodal approach with cultural, syntactic, and semantic feature projections to capture nuanced linguistic characteristics in Malayalam and Tamil.

4.2 byteSizedLLM

The team used a hybrid methodology, combining a customized BiLSTM network with a fine-tuned XLM-RoBERTa base model. The XLM-RoBERTa model was fine-tuned using Masked Language Modeling on a subset of the AI4Bharath dataset, enhancing its multilingual understanding. The dataset included original, fully transliterated, and partially transliterated data, allowing the model to learn robust cross-lingual representations and adapt to varying transliteration patterns. The BiLSTM layer fur-

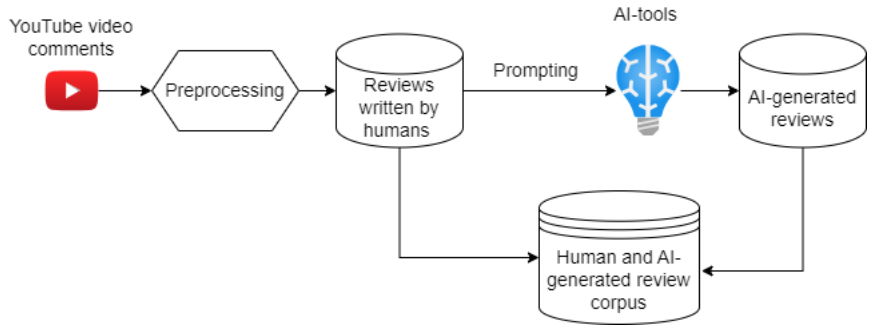

Figure 1: A block diagram explaining the process of creating the Malayalam dataset for the shared task.

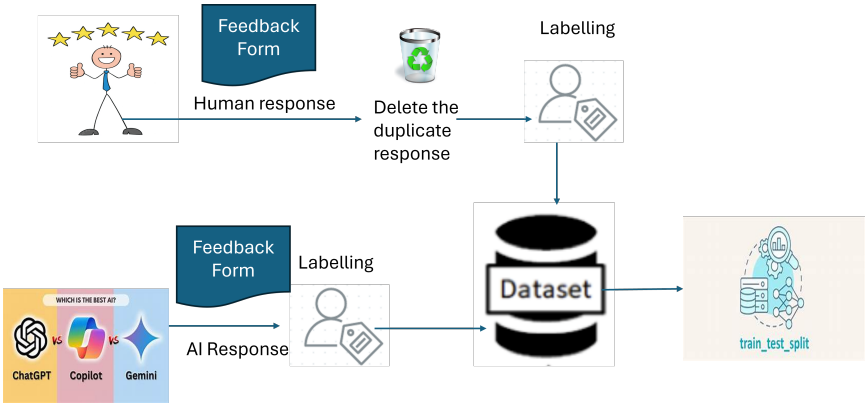

Figure 2: A block diagram explaining the process of creating the Tamil dataset for the shared task.

ther captures sequential dependencies, making it effective for multilingual tasks.

4.3 Girma

The team used Term Frequency-Inverse Document Frequency (TF-IDF) and logistic regression classifier for building the model. The other model proposed by this team had a BERT model trained using Dravidian language data, which outperformed the machine learning model trained using the TF-IDF features.

4.4 InnovateX

The proposed models address distinguishing AI-generated from human-written product reviews in Tamil and Malayalam using SVM, Logistic Regression (LR), and BERT-based transformers. Preprocessing involved text cleaning, tokenization, and label encoding. TF-IDF features (unigrams and bigrams) were used for SVM and LR, while BERT was fine-tuned for contextual understanding.

4.5 Cuet_Absolute_Zero

This team conducted experiments using machine learning models (random forest, support vector machine, decision tree and XGBoost), deep learning models (RNN, GRU, LSTM and BiLSTM) and transformer-based models. The authors increased the number of data points in each class by augmenting new data generated using backtranslation approach.

4.6 CODEGEEK

The model classifies Tamil text as AI-generated or human-generated using a combination of transformer embeddings and a Random Forest classifier. The model uses a pre-trained multilingual BERT model and tokenizer to generate high-dimensional numerical representations of the text's semantic and contextual meaning. These embeddings are used to train a Random Forest Classifier, which learns to classify text based on these embeddings. This combination of deep learning and traditional machine learning ensures effective classification for complex multilingual tasks.

4.7 CIC-NLP

They fine-tuned the XLM-RoBERTa model for text classification tasks in Malayalam and Tamil. Datasets were loaded, preprocessed, and tokenized

Team	F1-score	Rank
KaamKro	0.9199	1
Nitiz	0.9150	2
Three_Musketeers	0.9150	2
SSNTrio (J et al., 2025)	0.9147	3
byteSizedLLM (Kodali et al., 2025)	0.9000	4
Lowes	0.9000	4
CUET_NLP_FiniteInfinity (Hasan et al., 2025)	0.8999	5
TeamVision (S R et al., 2025)	0.8999	5
Necto (Dhasan, 2025)	0.8997	6
Cuet_Absolute_Zero-SIDRATUL	0.8996	7
MUNTAHA (Bijoy et al., 2025)		
Cuet_Absolute_Zero_run-Anindo Barua bijoy	0.8996	7
RATHAN (Thevakumar and Thevakumar, 2025)	0.8994	8
AnalysisArchitects (Jayaraman et al., 2025)	0.8850	9
CIC-NLP (Achamaleh et al., 2025)	0.8849	10
Girma (Bade et al., 2025)	0.8849	10
the_deathly_hallows (Shanmugavadivel et al., 2025)	0.8797	11
CODEGEEK	0.8748	12
MNLP	0.8550	13
AIstudent	0.8350	14
Friends	0.8298	15
Team_Risers (P et al., 2025)	0.8150	16
VKG	0.7834	17
AiMNLP (De and Vats, 2025)	0.7345	18
CUET_NetworkSociety (Aftahee et al., 2025)	0.7287	19
LinguAIsts	0.7100	20
NLP_goats (V K et al., 2025)	0.6849	21
KECLinguAIsts (Subramanian et al., 2025)	0.6697	22
InnovateX (A et al., 2025)	0.6449	23
powerrangers	0.6348	24
VRCLC	0.6310	25
SemanticCuetSync	0.5713	26
Miracle_makers	0.3333	27
HibiscusBots-CIOL	0.1299	28

Table 3: Ranklist of Malayalam sub-task

Team	F1-score	Rank
KEC_AI_NLP	0.9700	1
CUET_NLP_FiniteInfinity (Hasan et al., 2025)	0.9700	1
CIC-NLP (Achamaleh et al., 2025)	0.9600	2
KaamKro	0.9500	3
KEC-Elite-Analysts	0.9499	4
byteSizedLLM (Kodali et al., 2025)	0.9400	5
Nitiz - StarAtNyte	0.9300	6
VKG	0.9299	7
the_deathly_hallows (Shanmugavadivel et al., 2025)	0.9298	8
Team_Risers (P et al., 2025)	0.9197	9
NLP_goats (V K et al., 2025)	0.9099	10
Girma (Bade et al., 2025)	0.8998	11
Three_Musketeers	0.8900	12
AnalysisArchitects (Jayaraman et al., 2025)	0.8800	13
CODEGEEK	0.8678	14
InnovateX (A et al., 2025)	0.8600	15
KECLinguAIsts (Subramanian et al., 2025)	0.8598	16
RATHAN (Thevakumar and Thevakumar, 2025)	0.8368	17
CUET_NetworkSociety (Aftahee et al., 2025)	0.8182	18
AIstudent	0.8140	19
AiMNLP (De and Vats, 2025)	0.7287	20
Lowes	0.7083	21
powerrangers	0.6981	22
Friends	0.6834	23
HibiscusBots-CIOL	0.6745	24
Necto (Dhasan, 2025)	0.6745	24
LinguAIsts	0.6516	25
MNLP	0.6511	26
CUET-NLP_Big_O	0.6419	27
Cuet_Absolute_Zero_run - Anindo Barua bijoy	0.6311	28
Cuet_Absolute_Zero-SIDRATUL	0.6311	28
MUNTAHA (Bijoy et al., 2025)		
SSNTrio (J et al., 2025)	0.5989	29
TeamVision (S R et al., 2025)	0.5586	30
SemanticCuetSync	0.4857	31
Miracle_makers	0.3243	32

Table 4: Ranklist of Tamil sub-task

using the XLM-RoBERTa tokenizer, with labels mapped to numerical codes. The data was split into training and testing sets, and the model was trained using the Hugging Face Trainer API with parameters such as learning rate, batch size, and evaluation strategy. Post-training, we evaluated the model using metrics like accuracy, F1-score, confusion matrices, and ROC curves on the development dataset. Prediction CSV files for the test sets were saved for submission.

4.8 CUET_NetworkSociety

The team used a streamlined machine learning pipeline based on the DistilBERT model, which involved data preprocessing, tokenization, and model training. The preprocessing involved cleaning text to remove HTML tags, punctuation, and numbers, and normalizing whitespace. The tokenized text was then converted into a model-ready format using DistilBERT's tokenizer, ensuring maximum sequence length. The training phase involved splitting data into training and validation sets, with the model trained to maximize the F1 score.

4.9 KECLinguAIsts

The team preprocessed input data by cleaning, removing unwanted characters, and tokenizing reviews. They used TF-IDF vectorization to convert text into numerical features, capturing word importance in each language context. The training dataset was split into 80% for training and 20% for testing. For Tamil, they used Logistic Regression, Random Forest, and XG Boost, while for Malayalam, they used Logistic Regression, MNB, and SVM, each chosen for its ability to handle text data.

4.10 VKG

The proposed system employed the mBERT model (bert-base-multilingual-cased configuration) trained from scratch to classify Tamil and Malayalam product reviews as either human-written or AI-generated. For Tamil reviews, the model achieved a test accuracy of 98.77% and an F1-score of 0.99 for both classes after 5 epochs of training, demonstrating the potential of training multilingual models from scratch for this task, even with lim-

ited data. For Malayalam reviews, the model was trained for 8 epochs.

4.11 LinguAIsts

In this work, initially to preprocess the dataset, labels were encoded into binary values (1 for AI, 0 for humans). Each review was tokenized using BERT, and contextual embeddings were extracted using the [CLS] token, which captures the text's overall semantic meaning. A Support Vector Machine (SVM) classifier with a linear kernel used these embeddings as input features. 80% of the dataset was used to train the model, while the remaining 20% was used for evaluation.

4.12 Team_Risers

This team used a pre-trained language model fine-tuned specifically for Dravidian languages. This method involved preprocessing a custom dataset for compatibility with the model, which included text cleaning, tokenization, and encoding. The model was then trained on the dataset to adapt to the nuances of the Dravidian languages to correctly classify reviews as human-written or AI-generated.

4.13 Three_Musketeers

The team employed a combination of multilingual transformer models to classify AI-generated and human-generated text in both Malayalam and Tamil datasets. Specifically, for the Malayalam dataset, they utilized XLM-RoBERTa-Large, mBERT (Multilingual BERT), and IndicBERT to leverage their multilingual capabilities and contextual understanding of Indian languages. These models were fine-tuned on the dataset to optimize performance metrics such as F1-score and accuracy. For the Tamil dataset, they used mBERT due to its robust multilingual capabilities and proven effectiveness in handling diverse linguistic structures.

4.14 powerrangers

This team used K-Nearest Neighbor classifier for classifying the product reviews into human and ai categories.

4.15 NLP_goats

In this submission, they applied a machine learning approach for text classification by first preprocessing the text data, which involved removing punctuation, numbers, extra spaces, and converting text to lowercase. The processed text was then transformed into numerical features using TF-IDF with character-level n-grams (unigrams and bigrams), which helps capture important features, especially for languages such as Malayalam. A Logistic Regression classifier was trained on the transformed data, and the model's performance was evaluated using metrics such as accuracy, F1 score, precision, and recall, providing a comprehensive assessment of its effectiveness in classifying the text into predefined categories.

4.16 CUET-NLP_Big_O

The team utilized the BiLSTM+CNN model, which integrates convolutional and bidirectional recurrent layers for text classification. The model starts with an embedding layer with a vocabulary size of 10,000 and dimension of 128, then uses a Conv1D layer with 128 filters and a kernel size of 5. It captures contextual dependencies using a Bidirectional LSTM with 64 units per direction. The model refines features and ensures precise classification using a dense layer and softmax layer.

4.17 Rathan

The study used a pretrained multi-model ensemble approach for classification, using mT5-small, XLM-RoBERTa-base, Sentence-Transformers, and IndicBERTv2-MLM-only as feature extraction models. A dense neural network was trained on the extracted features for classification. A weighted averaging ensemble was used to combine the predictions from these models, with softmax probabilities weighted and averaged based on individual performances on the validation set. The final prediction was determined by selecting the class with the highest combined probability.

4.18 CIC-NLP

The team fine-tuned the XLM-RoBERTa model for text classification tasks in Malayalam and Tamil. Datasets were loaded, preprocessed, and tokenized using the XLM-RoBERTa tokenizer, with labels mapped to numerical codes. The data was split into training and testing sets, and the model was trained using the Hugging Face Trainer API with parameters such as learning rate, batch size, and evaluation strategy.

4.19 TeamVision

This team experimented with models like Bert, Naive Bayes, Random Forest, KNN, LSTM and Decision Tree combined with feature extraction

methods such as Bag of Words, TF-IDF, Count Vectorization, and n-grams. They identified BERT as the most accurate model for detecting AI-generated text in Tamil and Malayalam product reviews.

4.20 CUET_NLP_FiniteInfinity

This team employed Sarvam-1 and Gemma-2-2B, two advanced language models with capabilities in Tamil and Malayalam, among other languages.

4.21 HibiscusBots-CIOL

In this work, the team used language-specific models for each language to encode the text and obtain text embeddings. Additionally, they incorporated general Indic language embeddings to handle any cross-lingual nuances. These embeddings were then passed through a Multi-Layer Perceptron (MLP) for training, which facilitated sentiment prediction. They adopted an adaptive modeling approach, actively tracking the best model throughout the training process using the highest F1 score, and ultimately used the best-performing model for making predictions on the test data.

4.22 SSNTrio

In this work, the team upsampled teh data to eliminate class imbalance. BERT model was used for tokenization and used Tamil BERT and Malayalam BERT for classification.

4.23 AnalysisArchitects

They used SVM, IndiaBERT and ALBERT models to classify the task after encoding labels and vectorizing the text.

4.24 Lowes

This team used two BERT-based models such as multilingual BERT and L3Cube's monolingual BERT. In addition, then authors used a GPT-2 model with a causal language modeling (CLM) objective.

4.25 AiMNLP

This team proposed three models for this task: BERT embedding-based models, CNN+BiLSTM hybrid model and machine learning and machine learning ensemble models. BERT embedding-based models acheieved the highest performance score in both tasks.

4.26 KEC-Elite-Analysts

The team utilized a diverse set of models, including both traditional machine learning algorithms (e.g., Logistic Regression, Naive Bayes, SVM, Random Forest, Gradient Boosting) and deep learning approaches (e.g., HAN, DAN, mBERT, RoBERTa, ALBERT). This combination allowed us to compare and integrate the strengths of different techniques for effective classification.

4.27 Miracle_makers

This method leverages advanced NLP techniques, including preprocessing for Tamil text, feature extraction using embeddings (TF-IDF, GloVe, Word2Vec, BERT), and transfer learning with attention-based transformers (mBERT, RoBERTa). A fine-tuned binary classifier distinguishes AI-generated and human-written reviews, evaluated using metrics like accuracy, F1-score, and macro F1-score for robust detection.

4.28 The Deathly Hallows

In this work, the team implemented a deep learning-based approach to determine whether a given text was written by an AI or a human. They preprocessed the Tamil and Malayalam text data by normalizing, tokenizing, and removing stopwords to enhance feature extraction. For Tamil, Advertools was used to extract stopwords, while for Malayalam, they created a custom stopword list. After preprocessing, they used a pre-trained transformer model, such as BERT, to generate embeddings for the input text. These embeddings were then passed through a neural network for classification, where the model was trained to predict if the text was AI-generated or human-written.

4.29 SemanticCuetSync

This team used the Llama 3.2-3B model. At first they used a prompt to let the model know what to do. Then they finetuned the model with the training set. They used 10 epochs for Tamil and 45 epochs for Malayalam. Additionally, we quantized our Llama model to 4 bits. They used the model from Unsloth AI.

4.30 Friends

They used BERT for classification. This model leverages its powerful bidirectional contextual understanding to excel in tasks like natural language understanding and text classification. By incorporating BERT, this system effectively captures

semantic nuances, making it particularly adept at identifying subtle patterns and relationships in language data.

Different teams employed several methodologies to detect AI-generated reviews in Malayalam and Tamil. Transformer-based models, particularly BERT variants, dominated the submissions, leveraging their multilingual capabilities. Moreover, teams used hybrid architectures, such as combining BiLSTM with XLM-RoBERTa or CNN, to capture sequential and local patterns. Submissions based on traditional machine learning classifiers such as SVM, logistic regression, and random forest trained using TF-IDF features provide baselines. Teams used data augmentation techniques, such as back translation, to improve robustness. The models based on multilingual embeddings and ensemble strategies addressed linguistic nuances in the corpus. To summarize, teams built the models using both advanced deep learning and more traditional machine learning methods. However, transformer-based models excelled in identifying reviews generated using AI.

5 Conclusion

The shared task at DravidianLangTech@NAACL 2025 is organized to address the challenges of detecting AI-generated product reviews in Dravidian languages like Malayalam and Tamil. The models submitted by various teams to the shared task demonstrated the efficacy of transformer-based architectures in distinguishing human-written and AI-generated reviews, with top-performing teams achieving macro F1-scores exceeding 0.97 in Tamil and 0.91 in Malayalam. Hybrid models combining BiLSTM, CNN, or ensemble methods were effective in learning sequential and contextual information in the data. The performance of models trained using traditional machine learning classifiers with TF-IDF features lagged behind deep learning approaches, showing the significance of capturing more semantically rich embeddings. The novel dataset curated for this shared task is a significant contribution to the Dravidian language research. However, the artificial class balancing and lack of cross-domain generalization highlight the need for future work to incorporate more real-world characteristics in the data.

Acknowledgments

This work was conducted with the financial support from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2(Insight_2), supported in part of Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223.

References

Moogambigai A, Pandiarajan D, and Bharathi B. 2025. InnovateX@DravidianLangTech 2025: Detecting AI-Generated Product Reviews in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Tewodros Achamaleh, Abiola T.O, Lemlem Eyob, Mebiratu¸Mikiyas, and Grigori Sidorov. 2025. CIC-NLP @DravidianLangTech2025: Detecting AI-generated Product Reviews in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Sabik Aftahee, Tofayel Ahmmed Babu, MD Musa Kalimullah Ratul, Jawad Hossain, and Mohammed Moshiul Hoque. 2025. CUET_NetworkSociety@DravidianLangTech 2025: A Transformer-based Approach for Detecting AI-Generated Product Reviews in Low-Resource Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Girma Yohannis Bade, Muhammad Tayyab Zamir, Olga Kolesnikova, José Luis Oropeza, Grigori Sidorov, and Alexander Gelbukh. 2025. Girma@DravidianLangTech 2025: Detecting AI Generated Product Reviews. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Anindo Barua Bijoy, Sidratul Muntaha, Momtazul Arefin Labib, Samia Rahman, Udoy Das, and Hasan Murad. 2025. CUET_Absolute_Zero@DravidianLangTech 2025: Detecting Ai-Generated Product Reviews in Malayalam and Tamil Language Using Transformer Models. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Somsubhra De and Advait Vats. 2025. AiMNLP@DravidianLangTech2025: Unmask It! AI-Generated Product Review Detection in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language*

Technologies for Dravidian Languages. Association for Computational Linguistics.

Livin Nector Dhasan. 2025. Necto@DravidianLangTech: Fine-tuning Multilingual MiniLM for Text Classification in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Muhammad Fayaz, Atif Khan, Javid Ur Rahman, Abdullah Alharbi, M Irfan Uddin, and Bader Alouffi. 2020. Ensemble Machine Learning Model for Classification of Spam Product Reviews. *Complexity*, 2020(1):8857570.

Alessandro Gambetti and Qiwei Han. 2023. Dissecting AI-Generated Fake Reviews: Detection and Analysis of GPT-Based Restaurant Reviews on Social Media.

Md. Zahid Hasan, Safiul Alam Sarker, MD Musa Kalimullah Ratul, Kawsar Ahmed, and MohammedMoshiul Hoque. 2025. CUET_NLP_FiniteInfinity@DravidianLangTech 2025: Exploring Large Language Models for AI-Generated Product Review Classification in Malayalam. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bhuvana J, Mirnalinee T T, Rohan R, Diya Seshan, and Avaneesh Koushik. 2025. SSNTrio@DravidianLangTech 2025: Identification of AI Generated Content in Dravidian Languages using Transformers. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Sami Ben Jabeur, Hossein Ballouk, Wissal Ben Arfi, and Jean-Michel Sahut. 2023. Artificial Intelligence Applications in Fake Review Detection: Bibliometric Analysis and Future Avenues for Research. *Journal of Business Research*, 158:113631.

Abirami Jayaraman, Aruna Devi Shanmugam, Dharunika Sasikumar, and Bharathi B. 2025. AnalysisArchitects@DravidianLangTech 2025: BERT Based Approach For Detecting AI Generated Product Reviews In Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Rohith Gowtham Kodali, Durga Prasad Manukonda, and Maharajan Pannakkaran. 2025. AiMNLP@DravidianLangTech2025: Unmask It! AI-Generated Product Review Detection in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Minwoo Lee, Young Ho Song, Lin Li, Kyung Young Lee, and Sung-Byung Yang. 2022. Detecting Fake Reviews with Supervised Machine Learning Algorithms. *The Service Industries Journal*, 42(13-14):1101–1121.

Jiwei Luo, Guofang Nan, Dahui Li, and Yong Tan. 2023. AI-Generated Review Detection. *Available at SSRN 4610727*.

Sai Sathvik P, Muralidhar Palli, Keerthana NNL, Balasubramanian Palani, Jobin Jose, and Siranjeevi Rajamanickam. 2025. TeamRisers@DravidianLangTech 2025: AI-Generated Product Review Detection in Dravidian Languages Using Transformer-Based Embeddings. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Shankari S R, Sarumathi P, and Bharathi B. 2025. TeamVision@DravidianLangTech 2025: Detecting AI generated product reviews in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Malliga Subramanian, Vasantharan K, Prethish G A, and Vijayakumaran S. 2025. The_Deathly_Hallows@DravidianLangTech 2025: AI Content Detection in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Malliga Subramanian, Rojitha R, Mithun Chakravarthy Y, Renusri R V, and Kogilavani Shanmugavadivel. 2025. KECLinguAIsts@DravidianLangTech 2025: Detecting AI-generated Product Reviews in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Jubeerathan Thevakumar and Luheerathan Thevakumar. 2025. RATHAN@DravidianLangTech 2025: Annaparavai- Separate the Authentic Human Reviews from AI-generated one. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Shreeji Tiwari, Rohit Sharma, Rishabh Singh Sikarwar, Ghanshyam Prasad Dubey, Nidhi Bajpai, and Smriti Singhatiya. 2024. Detecting AI Generated Content: A Study of Methods and Applications. In *International Conference on Communication and Computational Technologies*, pages 161–176. Springer.

Srihari V K, Vijay Karthick Vaidyanathan, Mugilkrishna D U, and Durairaj Thenmozhi. 2025. NLP_goats@DravidianLangTech 2025: Detecting AI-Written Reviews for Consumer Trust. In *Proceedings of the Fifth Workshop on Speech, Vision, and*

Language Technologies for Dravidian Languages. Association for Computational Linguistics.

PS Venugopala, Amrith R Naik, Nidhish Shettigar, N Vaishnavi, Pranav R Bhat, and Pranesh Kumar Kodi. 2024. Identifying Deceptive AI Reviews: A Machine Learning Approach. In *2024 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*, pages 55–59. IEEE.

Mudasir Ahmad Wani, Mohammed ElAffendi, and Kashish Ara Shakil. 2024. AI-Generated Spam Review Detection Framework with Deep Learning Algorithms and Natural Language Processing. *Computers (2073-431X)*, 13(10).

Konstantinos F Xylogiannopoulos, Petros Xanthopoulos, Panagiotis Karampelas, and Georgios A Bakamitsos. 2024. ChatGPT Paraphrased Product Reviews Can Confuse Consumers and Undermine Their Trust in Genuine Reviews. Can You Tell the Difference? *Information Processing & Management*, 61(6):103842.