CUET_NetworkSociety@DravidianLangTech 2025: A Transformer-based Approach for Detecting AI-Generated Product Reviews in Low-Resource Dravidian Languages

Sabik Aftahee*, **Tofayel Ahmmed Babu***, **MD Musa Kalimullah Ratul***
Jawad Hossain and Mohammed Moshiul Hoque
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
{u1904005, u1904024, u1904071, u1704039}@student.cuet.ac.bd
moshiul_240@cuet.ac.bd

Abstract

E-commerce platforms face growing challenges regarding both consumer trust and review authenticity because of the growing number of AI-generated product reviews. Low-resource languages such as Tamil and Malayalam face limited investigation by AI detection techniques because these languages experience constraints from sparse data sources and complex linguistic structures. The research team at CUET_NetworkSociety took part in the AI-Generated Review Detection contest during the DravidianLangTech@NAACL 2025 event to fill this knowledge void. Using a combination of machine learning, deep learning, and transformer-based models, we detected AI-generated and human-written reviews in both Tamil and Malayalam. Among the approaches used, DistilBERT was found to be better suited to detect AI-Generated Reviews, which underwent an advanced preprocessing pipeline and hyperparameter optimization using the Transformers library. This approach achieved a Macro F1-score of 0.81 for Tamil (Subtask 1), securing 18^{th} place, and a score of 0.72 for Malayalam (Subtask 2), ranking 25^{th}.

1 Introduction

Online authenticity and reliability face serious obstacles because of the recent growth of AI-generated content in the current era. Product reviews experience direct negative impacts because customers heavily depend on them during purchasing decisions. These reviews are being generated by AI, often mimicking human reviews. This rise in AI-generated reviews has far-reaching implications, as it undermines trust in online marketplaces, misleads consumers, and distorts market dynamics (Raja et al., 2023). Thus, the need to detect those contents is very imminent. Researchers created solid detection methods for AI-generated content in different languages through advanced deployments of natural language processing and machine learning systems (LekshmiAmmal et al., 2022). However, most studies have focused mainly on high-resource languages such as English and Spanish, leaving low-resource languages such as Tamil and Malayalam underrepresented (Hegde and Shashirekha, 2021). The Dravidian language family poses distinctive detection challenges because of its complex morphological structure combined with semantic richness and various dialectal variations (Coelho et al., 2023). Research on AI-generated review detection in Tamil and Malayalam languages faces negligible attention despite their economic relevance due to the limited available datasets and the intricate linguistic structures of these languages (Krishnan et al., 2024). Online review credibility and user trust in the specified regions become essential to remedy. The research establishes a reliable method for detecting product evaluations created by AI in Tamil and Malayalam by focusing on this specific problem. The system explored various machine learning (ML), deep learning (DL), and transformer-based models to overcome the linguistic challenges of detecting AI-generated product reviews in Dravidian languages. The critical contributions of this work are:

- Investigated several ML, DL, and transformer-based models to detect AI-generated reviews in Tamil and Malayalam.
- Evaluated the performance of employed models and provided a comparative analysis to identify the most effective approach for detecting AI-generated content in these Dravidian languages.

2 Related Work

Recent research on fake review detection has focused on using machine learning (ML) and deep learning (DL) techniques to tackle this problem

*Authors contributed equally to this work.

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 600–606
May 3, 2025 ©2025 Association for Computational Linguistics

in various datasets. Barbado et al. (2019) proposed the Fake Feature Framework (F3) to detect fake reviews using user-centric and review-centric features, working with a custom Yelp consumer electronics dataset and the DOSA dataset, achieving an F1-score of 82% with AdaBoost. Raheem and Chong (2024) compared deep learning models (LSTM, CNN, hybrid) with transformers (DistilBERT) for fake review detection, utilizing the Yelp Reviews dataset, and found DistilBERT achieved 96% accuracy. Abd-Alhalem et al. (2024) integrated deep learning with aspect-based sentiment analysis using the OSF dataset, achieving 97.73% accuracy with an LSTM-based model. Vashist et al. (2024a) employed ensemble machine learning techniques (XGBoost, Random Forest) combined with BERT for detecting fake reviews on the OSFHOME dataset, achieving 98.2% accuracy with BERT. Deshai and Bhaskara Rao (2023) explored hybrid deep learning models (CNN-LSTM and LSTM-RNN) with GloVe embeddings for fake review and rating detection on Amazon Unlocked Mobile and Hotel datasets, reaching 93.07% accuracy. Ennaouri and Zellou (2023) reviewed various ML techniques and ensemble voting for fake review detection, reporting 97.5% accuracy on the CloudArmor dataset. Saini and Khatarkar (2023) analyzed fake news detection methods, which can be applied to fake reviews, using the WELFake dataset, achieving 96.73% accuracy. Veda et al. (2024) proposed a hybrid model combining BERT embeddings and ensemble methods (Random Forest, XGBoost) on the Public Fake Reviews dataset, achieving 86.45% accuracy with a stacking classifier. Rajesh et al. (2023) utilized sentiment analysis with traditional ML classifiers on Amazon Reviews, achieving 85% accuracy with Logistic Regression and Count Vectorizer. Wagh et al. (2024) applied Random Forest and NLP techniques on the Amazon Yelp Academic dataset to detect spam reviews, achieving 89.49% accuracy. Vashist et al. (2024b) used CNN and SVM models for detecting fraudulent reviews on a custom dataset, with CNN achieving 89% accuracy. V et al. (2023) provided a general overview of ML techniques for fake review detection but did not specify a dataset or accuracy. Alkomah and Sheldon (2023) reviewed advancements in fake news detection techniques, which could be adapted for fake review detection, but did not provide specific performance metrics. Sharma et al. (2023) explored hybrid deep learning models, integrating Bi-LSTM and CNN for fake review detection, highlighting the effectiveness of combining contextual and sequential information, with the highest accuracy reported at 95%. Transformers have revolutionized AI-generated content detection with models like BERT, RoBERTa, and their multilingual variants. LekshmiAmmal et al. (2022) demonstrated the potential of transformers in detecting toxic spans in Tamil, while Bafna et al. (2023) developed a RoBERTa-BiLSTM hybrid model that achieved a significant boost in accuracy for AI-generated text detection. Moreover, Coelho et al. (2023) focused on Malayalam, showcasing the efficacy of TF-IDF combined with ensemble ML models for detecting fake reviews in low-resource languages, achieving a macro F1 score of 0.831.

3 Task and Dataset Descriptions

For the goal of detecting AI-generated product reviews in Tamil and Malayalam languages, we utilized datasets specifically curated for this task (Premjith et al., 2025). The datasets consist of training, validation, and test data with detailed distributions as outlined below.

3.1 Tamil Dataset

The Tamil data set comprises a balanced distribution of human-generated and AI-generated reviews. Table 1 presents the statistics of the Tamil dataset in the training, validation, and test sets.

Classes	**Train**	**Test**	W_t	U_w
AI	405	48	3583	1423
Human	403	52	2428	1281
Total	808	100	6011	2704

Table 1: Class-wise distribution of training and test sets for Tamil, where W_T denotes the total number of words, and U_W denotes the number of unique words

3.2 Malayalam Dataset

The Malayalam dataset also maintains a balanced distribution across AI-generated and human-written reviews. Detailed statistics are shown in Appendix A. Both datasets provide a nearly equal class distribution across AI-generated and human-written reviews, ensuring a balanced evaluation setup. The unique word counts highlight the linguistic diversity and vocabulary range in both Tamil and Malayalam datasets. The implementa-

tion details are available at the link[1].

4 Methodology

The following section details a complete set of procedures along with methodologies which tackle the existing text classification hurdles described earlier. Figure 1 illustrates the abstract process of detecting AI-generated reviews. Our method uses machine learning alongside deep learning and transformer-based models while optimizing and tuning these models to boost their performance in text classification applications.

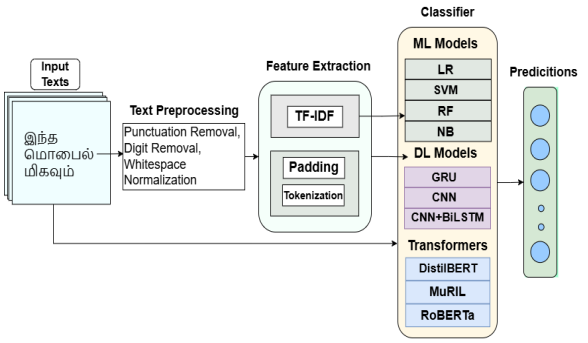

Figure 1: Abstract process of AI Generated Review detection

4.1 Pre-processing and Feature Extraction

We used extensive pre-processing techniques to normalize input data which created essential conditions for model training success. The preprocessing step involved removal of tags, punctuation, and numbers to ensure uniformity. Additionally, the text data was transformed using TF-IDF vectorization, which has been shown to effectively identify important words within a document.

4.2 ML Models

We used baseline machine learning models such as LR, SVM, RF, and Naive Bayes for the first assessments. Multiple metrics including accuracy, precision, recall, and F1 score were used for evaluation to measure model performance. TF-IDF vectorization was implemented which transformed text data into its top 1000 terms for efficient classification. Logistic regression received configuration adjustments for better convergence performance by setting its maximum iteration threshold to 1000.

[1] https://github.com/pr0ximaCent/DravidianLangtech-2025

4.3 DL Models

Additionally, a combination of deep learning architectures such as CNNs and BiLSTM networks were used to further assess the work. These models are particularly better at capturing complex patterns in text data, making them suitable for the nuanced demands of natural language understanding. The training of these models was systematically conducted, involving the tuning of hyperparameters like the number of layers, dropout rates, and learning rates to optimize performance. These models utilized embeddings of dimension 128 and were optimized with the Adam optimizer at a learning rate of 1e-3, training on batches of 32 samples. The classification output was derived using a sigmoid activation function.

4.4 Transformer Models

The highlight of our methodology was the application of transformer-based models to detect AI-generated product reviews in Dravidian languages, celebrated for their efficiency and robustness in handling various NLP tasks (Fariello et al., 2024). We fine-tuned three transformer-based models (MuRIL-BERT, RoBERTa and DistilBERT) on our dataset, which entailed several pivotal steps: text data tokenization using the transformers library tokenizer, integration of early stopping, and dynamic learning rate adjustments to forestall overfitting while expediting convergence to an optimal model state. Training was meticulously executed using Hugging Face's `Trainer` API, incorporating strategies such as batch size optimization and validation-based tuning to ensure the model's effectiveness (Forte and Marotta, 2024; Raja and Wani, 2023). The models were trained for up to four epochs, with periodic evaluations to adjust training parameters based on real-time performance metrics. Each model was validated with a distinct set to ensure better reliability & generalization on unseen data (Chaka, 2024; Ara et al., 2024).

5 Result Analysis

We observed the ability of Machine Learning (ML) and Deep Learning (DL) with Transformer-based models to identify AI-generated write-ups from authentic human reviews within Tamil and Malayalam datasets. The measurement of classifier performance included precision (P), recall (R), F1-score (F1), and accuracy (A). A comprehensive summary of model performance is presented in

Table 2.

Classifier	P	R	F1	A
Malayalam				
Logistic Regression (LR)	0.58	0.60	0.59	0.61
SVM	0.55	0.56	0.555	0.57
Random Forest (RF)	0.53	0.53	0.53	0.54
Naive Bayes (NB)	0.50	0.48	0.49	0.51
CNN	0.63	0.64	0.635	0.65
GRU	0.66	0.66	0.66	0.67
CNN-LSTM	0.65	0.65	0.65	0.66
CNN-BiLSTM	0.67	0.68	0.67	0.69
MuRIL-BERT	0.68	0.68	0.68	0.69
RoBERTa	0.67	0.67	0.67	0.68
DistilBERT	**0.75**	**0.71**	**0.72**	**0.75**
Tamil				
Logistic Regression (LR)	0.60	0.62	0.61	0.63
SVM	0.57	0.58	0.57	0.59
Random Forest (RF)	0.55	0.55	0.55	0.56
Naive Bayes (NB)	0.52	0.50	0.51	0.53
CNN	0.65	0.66	0.655	0.67
GRU	0.68	0.68	0.68	0.69
CNN-LSTM	0.67	0.67	0.67	0.68
CNN-BiLSTM	0.69	0.70	0.69	0.71
MuRIL-BERT	0.70	0.70	0.70	0.71
RoBERTa	0.69	0.69	0.69	0.70
DistilBERT	**0.85**	**0.78**	**0.81**	**0.76**

Table 2: Performance Comparison of Classifiers Across ML, DL, and Transformer Models for Tamil and Malayalam

Logistic Regression (LR) achieved F1 scores of 0.59 for Malayalam and 0.61 for Tamil, highlighting its effectiveness in modeling linear relationships. However, traditional machine learning models like SVM, Random Forest (RF), and Naive Bayes (NB) performed weaker, with F1 scores between 0.49 and 0.56, showing limitations in capturing complex semantic patterns in text. These models rely heavily on manual feature engineering and may struggle with high-dimensional data like text. Their inability to automatically capture semantic and syntactic complexities resulted in a subpar performance in language tasks.

The CNN-BiLSTM model achieved F1 scores of 0.67 for Malayalam and 0.69 for Tamil, effectively capturing complex relationships and long-range dependencies. Alternative deep learning models like GRU and CNN-LSTM performed competitively, with F1 scores between 0.66 and 0.68. Deep learning architectures can learn hierarchical representations of data, capturing local and global text dependencies. This ability enabled them to understand context and semantics better than traditional models, leading to improved performance.

Transformer-based models, especially DistilBERT, outperformed traditional ML and deep learning models, achieving the highest F1 scores (0.75 for Malayalam, 0.81 for Tamil) using dynamic contextual embeddings. MuRIL-BERT and RoBERTa followed closely with F1 scores between 0.68 and 0.70, showcasing strong performance, particularly for low-resource languages. Traditional ML methods and CNN-BiLSTM demonstrated adequate performance, but transformers surpassed them to achieve better results. Transformers utilize self-attention mechanisms to effectively weigh the importance of different words in a sentence. This capability enabled them to capture complex patterns and contextual relationships more efficiently than previous architectures, leading to superior performance in language understanding tasks. Hyperparameter tunings have been discussed in Appendix B.

6 Error Analysis

An in-depth error analysis was conducted using both quantitative and qualitative methods to evaluate the performance of the proposed model.

6.1 Quantitative Error Analysis

To further understand the performance of the models, a quantitative analysis was performed using confusion matrices for Tamil and Malayalam. Figures 2 and 3 illustrate the confusion matrices for both languages.

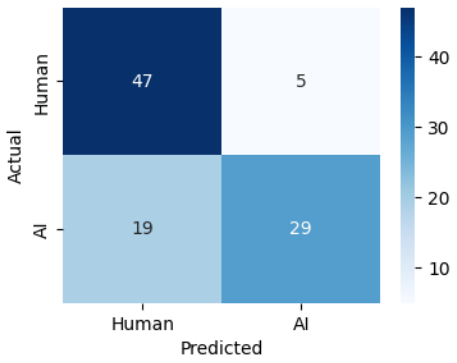

Figure 2: Confusion Matrix for Tamil Dataset

The Tamil confusion matrix shows 47 correctly identified human reviews, with 5 misclassified as AI. However, 19 AI reviews were wrongly labeled as human, while 29 were correctly classified. Similarly, in Malayalam, 79 human reviews were correctly identified, but 21 were misclassified. The model also correctly predicted 74 AI reviews, though 26 were mistaken for human, highlighting challenges in AI text detection.

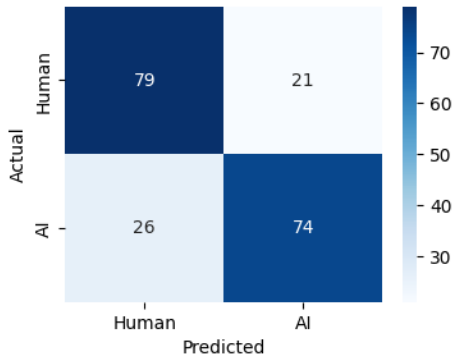

Figure 3: Confusion Matrix for Malayalam Dataset

Text Sample	Actual	Predicted
അടച്ച പൈസ നഷ്ടപ്പെടാതെ കിട്ടണമെങ്കിൽ എൽ.ഐ.സി മാത്രേ ഒള്ളൂ. 100% സോവെറിക്ൻ ഗുയാരണ്ടീ.	Human	Human
കോവയ്ക്ക ഉപ്പിലിട്ത് ഞാൻ ഇതുവരെ കഴിച്ചിട്ടില്ല, കഴിക്കാൻ മനസില്ല.	AI	Human
ഞാൻ 19227 മന്തിലി ടാറ്റ അടക്കുന്നു ഫോർ 5 വർഷം. മെച്യൂരിറ്റി ര്ം വർഷം . ഏകദേശം 10 കോടിക്ക് മുകളിൽ റിട്ടേൺ കിട്ടും	Human	Human
പഴകിയ മീനും കറികളും കഴിച്ചു പല പ്രാവശ്യം ഫുഡ് പൊയ്സണിംഗിനെ നേരിട്ടിട്ടുണ്ട്.	AI	Human
കോവയ്ക്ക ഉപ്പിലിട്ത് ഞാൻ ഇതുവരെ കഴിച്ചിട്ടില്ല, കഴിക്കാൻ മനസില്ല.	AI	AI

Figure 5: Few examples of predicted outputs by the proposed (DistilBERT) model for Malayalam

Some errors in quantitative analysis, as reflected in confusion matrices, stem from overlapping linguistic patterns between AI-generated and human-written text. The model struggles to differentiate when AI text mimics human fluency or human reviews exhibit generic phrasing, leading to misclassifications.

6.2 Qualitative Error Analysis

To complement the quantitative analysis, a qualitative examination of the misclassified examples was conducted. Figures 4 and 5 present a few representative examples of predicted outputs by the model.

Text Sample	Actual	Predicted
என் ஹெட்செட்/ஈர்பாட் பிராண்ட் பயன்படுத்தியபோது, சில நேரங்களில் ஆடியோ குவாலிட்டி சரியில்லாமல், சத்தம் மிகவும் குறைந்துவிடுகிறது.	AI	AI
நான் அண்மையில் வாங்கிய ஒரு குக்கர் ஆரோக்கிய உணவு தயாரிப்பதற்காக சரியான தேர்வாக இல்லை.	AI	Human
அணிவதற்கு நன்றாக இருக்கும்	Human	Human
அதிக வாசனை வாந்தி	Human	AI
அதிகமக பயன்பதுதினல் தலை சுட்ரும்	Human	Human

Figure 4: A few examples of predicted outputs by the proposed (DistilBERT) model for Tamil.

The qualitative analysis revealed misclassifications where AI reviews were labeled as human-written due to colloquial language, and human reviews were mistaken for AI-generated due to generic phrasing. These errors highlight the need for improved contextual differentiation and feature extraction to reduce misclassifications.

7 Conclusion

This research explored the detection of AI-generated product reviews in low-resource Dravidian languages, specifically Tamil and Malayalam, using machine learning, deep learning, and transformer-based models. The study found that traditional ML models like Logistic Regression and SVM struggled to capture the intricate linguistic features of these languages. Deep learning approaches, such as CNN-BiLSTM, improved performance by better modeling text dependencies. However, the transformer-based DistilBERT model demonstrated the highest effectiveness, achieving the best F1-scores for both Tamil and Malayalam datasets. The research outcome confirms that transformer models demonstrate high capability when used for text classification in languages with minimal resources. The next step should concentrate on using extensive datasets as well as better fine-tuning methods and contextual elements to enhance the accuracy rate.

Limitations

Despite the contributions in detecting AI-generated reviews in Tamil and Malayalam, several limitations remain: (i) Pre-trained transformer models like DistilBERT may be limited by their training corpus, affecting their ability to capture the nuances of these languages. (ii) The small datasets used constrained the models' generalization to unseen data. (iii) Linguistic complexities, such as code-mixing and dialect variations, present challenges in accurate text classification.

Acknowledgments

We thank the DravidianLangTech 2025 shared task organizers for running this task. This work was supported by the Directorate of Research & Extension (DRE), Chittagong University of Engineering & Technology (CUET).

References

Samia M. Abd-Alhalem, Hesham A. Ali, Naglaa F. Soliman, Abeer D. Algarni, and Hanaa Salem Marie. 2024. Advancing e-commerce authenticity: A novel fusion approach based on deep learning and aspect features for detecting false reviews. *IEEE Access*, 12:1–17.

Bushra Alkomah and Frederick Sheldon. 2023. Advancements in fake news detection using machine and deep learning models: Comprehensive literature review. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1–6.

Anjuman Ara, Md Sajadul Alam, Kamrujjaman, and Afia Farjana Mifa. 2024. A comparative review of ai-generated image detection across social media platforms. *Global Mainstream Journal of Innovation, Engineering Emerging Technology*, 3(1):11–19.

R. Bafna, M. Jain, and P. Sharma. 2023. Roberta and bilstm hybrid architecture for ai-generated text detection. In *Proceedings of the 2023 International Conference on Natural Language Processing*, pages 233–241.

Rodrigo Barbado, Oscar Araque, and Carlos A. Iglesias. 2019. A framework for fake review detection in online consumer electronics retailers. *Information Processing Management*, 56(4):1234–1244.

C. Chaka. 2024. Differentiating between ai-generated and human-written text using ai detection tools. *Journal of Applied Learning and Teaching*, 7(1):118–126.

Sharal Coelho, Asha Hegde, and Hosahalli Lakshmaiah Shashirekha. 2023. Mucs@dravidianlangtech2023: Malayalam fake news detection using machine learning approach. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 288–292.

N Deshai and B Bhaskara Rao. 2023. Deep learning hybrid approaches to detect fake reviews and ratings. *Journal of Scientific & Industrial Research*, 82:120–127.

Mohammed Ennaouri and Ahmed Zellou. 2023. Machine learning approaches for fake reviews detection: A systematic literature review. *Journal of Web Engineering*, 22:821–848.

Serena Fariello, Giuseppe Fenza, Flavia Forte, Mariacristina Gallo, and Martina Marotta. 2024. Distinguishing human from machine: A review of advances and challenges in ai-generated text detection. *International Journal of Interactive Multimedia and Artificial Intelligence*, 8(5):1–12.

F. Forte and M. Marotta. 2024. Machine learning models for ai-generated text analysis. *Computational Intelligence Review*, 14(3):234–249.

Asha Hegde and Hosahalli Lakshmaiah Shashirekha. 2021. Urdu fake news detection using ensemble of machine learning models. *CEUR Workshop Proceedings*, pages 132–141.

S. Krishnan, R. Babu, and P. Nair. 2024. Detecting ai-generated text: A study on the performance of ml and dl models in dravidian languages. *Journal of Artificial Intelligence Research*, 17(4):254–271.

Hariharan LekshmiAmmal, Manikandan Ravikiran, and Anand Kumar Madasamy. 2022. Nitk-it nlp@tamilnlp-acl2022: Transformer based model for toxic span identification in tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 75–78, Dublin, Ireland. Association for Computational Linguistics.

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.

Mafas Raheem and Yi Chien Chong. 2024. E-commerce fake reviews detection using lstm with word2vec embedding. *CIT.2024*, 100:70–80.

K. Raja and S. Wani. 2023. Multilingual sentiment analysis for fake review detection. *International Journal of Computational Linguistics*, 12(1):88–102.

R. Raja, A. Kumar, and S. Joseph. 2023. Fake news detection in low-resource languages: Challenges and advancements. *Computational Linguistics Review*, 15(2):123–137.

N Rajesh, AC Ramachandra, Ayush Tomar, Heman Kumawat, Anurag Prasad, and Ramprasad Poojary. 2023. Fake reviews detection based on sentiment analysis using ml classifiers. *IEEE International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE)*.

Parul Saini and Virendra Khatarkar. 2023. A review on fake news detection using machine learning. *SMART MOVES JOURNAL IDSCIENCE*, 9:6–9.

A. Sharma, S. Kumar, and R. Gupta. 2023. Evaluating convolutional neural networks for text classification tasks in low-resource languages. *Journal of Computational Linguistics*, 49(2):123–135.

Arpitha S V, Ashwitha H N Jois, Bhargavi V M, Deeksha A H, and Sreedevi S. 2023. Detecting fake reviews in e-commerce platform. *International Journal of Advanced Research in Computer and Communication Engineering*, 12:1–6.

Ansh Vashist, Arul Keswani, Varda Pareek, and Tarun Jain. 2024a. Detecting fake reviews on e-commerce platforms using machine learning. In *2024 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS)*, pages 1–6.

Ansh Vashist, Arul Keswani, Varda Pareek, and Tarun Jain. 2024b. Detecting fraudulent reviews in e-commerce platforms. *2024 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS)*, pages 1–6.

Chitti Reddy Veda, Muni Sekhar Velpuru, Namburu Apoorva, Hammikolla Akshaya, N Vishnu, and Sai Prakhyath Siripuram. 2024. Fake review identification using hybrid fusion of machine learning and natural language processing techniques. *IEEE Access*.

Yogansh Wagh, Sarfaraz Ali, Apurva Bobade, Aditi Bhalekar, and Rajaram Ambole. 2024. E-commerce spam review detection using machine learning. *Journal of Information Systems and Renewable Energy Management (JISREM)*, pages 1–4.

A Class-wise Distribution of Malayalam Dataset

Table A.1 shows class-wise distribution of training and test sets for the Malayalam language, including the number of total and unique words in each category. The dataset is divided into AI-generated and Human-written texts, with an equal split between training and test samples. The statistics, such as the total words (W_T) and unique words (U_W), highlight the lexical diversity within each class. This information is crucial for understanding the dataset composition and its impact on model training and evaluation.

Classes	**Train**	**Test**	W_t	U_w
AI	400	100	5174	3138
Human	400	100	8201	4819
Total	800	200	13375	7957

Table A.1: Class-wise distribution of training and test sets for Malayalam where W_T and U_W, denotes total and unique words respectively .

B Tuned Hyperparameters

Table B.1 shows the fine-tuned hyperparameters for AI vs. Human text classification tasks using DistilBERT. A learning rate of 5×10^{-5} ensures stable convergence, while a batch size of 16 balances memory and training stability. The model trains for 4 epochs with a max sequence length of 256 to capture longer texts efficiently. Cross-entropy loss is used for classification, with AdamW as the optimizer and a weight decay of 0.01 to prevent overfitting. Gradient accumulation steps (2) simulate a larger batch size of 32, while 300 warmup steps and a linear learning rate scheduler help stabilize training. These hyperparameters were fine-tuned to maximize accuracy while maintaining computational efficiency.

Hyperparameter	**Value**
Learning Rate	5×10^{-5}
Per Device Batch Size	16
Number of Epochs	4
Max Sequence Length	256
Loss Function	Cross-Entropy Loss
Optimizer	AdamW
Weight Decay	0.01
Gradient Accumulation Steps	2
Warmup Steps	300
Learning Rate Scheduler	Linear

Table B.1: Tuned hyperparameters used for the AI vs. Human text classification task using DistilBERT.