LexiLogic@DravidianLangTech 2025: Detecting Fake News in Malayalam and AI-Generated Product Reviews in Tamil and Malayalam

Souvik Bhattacharyya*, **Pranav Gupta***, **Niranjan Kumar M, Billodal Roy**
Lowe's
Correspondence: {souvik.bhattacharyya, pranav.gupta, niranjan.k.m, billodal.roy}@lowes.com

Abstract

Fake news and hard-to-detect AI-generated content are pressing issues in online media, which are expected to exacerbate due to the recent advances in generative AI. Moreover, tools to keep such content under check are less accurate for languages with less available online data. In this paper, we describe our submissions to two shared tasks at the NAACL Dravidian Language Tech workshop, namely detecting fake news in Malayalam and detecting AI-generated product reviews in Malayalam and Tamil. We obtained test macro F1 scores of 0.29 and 0.82 in the multi-class and binary classification subtasks within the Malayalam fake news task, and test macro F1 scores of 0.9 and 0.646 in the task of detecting AI-generated product reviews in Malayalam and Tamil respectively.

1 Introduction

The proliferation of AI-generated content and misleading content such as fake news has resulted in concerns in multiple domains such as e-commerce, news and social media, and other digital domains. Existing tools to detect such content have been restricted to high-resource languages. Moreover, it is challenging for language models to learn complex and rapidly evolving trends and cultural attributes that determine whether something is false or misleading.

With the advent of large language models (LLMs) like GPT (Radford et al., 2018) and Llama (Touvron et al., 2023), generating machine-produced product reviews has become easier than ever. This has huge potential of misleading consumers into purchasing items they might not otherwise choose. As AI tools continue to advance, distinguishing between machine-generated and human-authored text is becoming increasingly challenging. Watermarking LLM output is a novel approach to mitigating the spread of LLM-generated text on the internet, whether in the form of product reviews, fake news, or propaganda on social media. However, due to the lack of consensus among major corporations, ethical concerns, and the availability of open-weight models, watermarking is no longer a consistently viable solution. This underscores the need to explore alternative approaches to address the growing challenge of detecting machine-generated content.

In this paper, we describe our submissions to 2 tasks at the Dravidian Language Tech workshop at NAACL 2025, namely fake news detection in Malayalam and detecting AI-generated product reviews in Tamil and Malayalam. The fake news task had 2 sub-tasks: performing binary classification of textual news as "original" or "fake," and performing multi-class classification of textual news as "half true," "partly false," "mostly false," and "false". On the test dataset, our submissions ranked **6th** out of 16th and **12th** out of 21 submissions in the binary and multi-class classification sub-tasks respectively.

The task on detecting AI-generated product reviews in Tamil and Malayalam consisted of product reviews with binary labels "human" (human generated) and "AI" (AI-generated). Our submissions ranked **4th** among 51 teams and **30th** among 54 teams in Malayalam and Tamil respectively.[1]

2 Related Work

While fake news detection efforts in Malayalam have been limited, several studies have addressed fake news detection in social media platforms using deep learning techniques (Shu et al., 2017; Ghosh and Mitra, 2017; Dhar and Agarwal, 2018; Subramanian et al., 2025). For low-resource languages like Malayalam, approaches such as transfer learning by fine-tuning multilingual BERT based mod-

*These authors contributed equally to this work.

[1]The code for this work is available at https://github.com/prannerta100/naacl2025-dravidianlangtech.

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 526–531

els (Devlin et al., 2019a; Dabre et al., 2022) have shown promise in earlier shared tasks.

As LLMs become more prevalent, research on detecting AI-generated content has grown. GPTZero (Habibzadeh, 2023) is one such tool developed to address concerns about academic plagiarism, using metrics like perplexity and burstiness, though it has been criticized for its false positive rate. Luo et al., 2023 introduced a supervised learning approach for detecting AI-generated reviews by categorizing linguistic features and training classifiers like kNN, AdaBoost, and SVM. Studies by Kirchenbauer et al., 2024 have focused on watermarking LLM outputs by embedding statistical patterns into machine-generated text that can still be detected algorithmically despite alterations like token replacements or paraphrasing. In contrast, DetectGPT, proposed by Mitchell et al., 2023, avoids the need for a separate classifier or explicit watermarking. Instead, it calculates *perturbation discrepancy* using log probabilities from the model of interest and random perturbations applied to the passage, checking if this discrepancy exceeds a predefined threshold. More recent work by Bahad et al., 2024 adopts OpenAI's approach of fine-tuning a RoBERTa-based model (Liu et al., 2019) on a diverse dataset which demonstrated strong performance in identifying the source language model among multiple candidates.

3 Fake News Detection

3.1 Binary Classification

In this task, we were given a dataset (Subramanian et al., 2024, 2023; Devika et al., 2024) of news in Malayalam, with labels "original" and "fake." The dataset consisted of social media posts in pure and code-mixed Malayalam, both in English and Malayalam scripts.

The details of the dataset are given in Table 1

Label	Train set	Dev set	Test set
Original	1658	409	512
Fake	1599	406	507

Table 1: Dataset details for the binary classification subtask

We tried the following 4 models for this subtask:

1. **TFIDF + Logistic Regression:** TFIDF vectorization was a popular method for creating features out of textual data before the advent of foundational neural language models such as BERT. Moreover, logistic regression on top of TFIDF features provides a simple, linear baseline with lesser chances of overfitting. We used the default 'scikit-learn' parameters for training the binary classifier, with a maximum solver iteration parameter of 100.

2. **Fasttext:** Fasttext (Joulin et al., 2016) is a library for efficient learning of word representations and text classification. It uses shallow neural networks for text classification, and includes other in-built optimizations for efficient model training.

3. **GPT-4o:** GPT-4o is an instruction-finetuned large language model by OpenAI, used for a variety of NLP applications. We used GPT-4o with selected training examples and the following prompt: *(system) You are an NLP expert helping classify Malayalam fake news. Before outputting, you will think what the text means within the cultural context of a Malayalam speaker.*
(user) You are a classifier. Use the training data below to classify each text as 'original' or 'fake', output only a json that is a list of records with fields 'text' and 'prediction':

4. **Malayalam BERT**: Malayalam BERT is a monolingual BERT model trained from publicly available monolingual Malayalam datasets (Joshi, 2022). We finetuned this BERT model for the binary classification subtask, given the model's ability to understand Malayalam text. Larger multilingual models are harder to finetune, hence we can expect Malayalam BERT to capture the nuances of Malayalam fake news better. We trained the model for 5 epochs on the train dataset, while using a learning rate of 2×10^{-5}, a weight decay regularization parameter of 0.01, and a per-device batch size of 32.

Table 2 summarizes the performance of our models we tried for this sub-task. We see that Malayalam BERT outperforms the other models. While Fasttext achieves a similar test accuracy as Malayalam BERT, the train-test performance gap is much higher, indicating overfitting to the train set. Such overfitting is not observed in the finetuned Malayalam BERT.

Model	Train F1	Test F1
TFIDF + Log. Reg.	0.928	0.769
FastText	0.9957	0.805
GPT-4o	-	0.782
`malayalam-bert`	0.851	**0.808**

Table 2: Train and Test set Macro F1 scores for the binary classification sub-task

3.2 Multi-class Classification

In this task, we were given a dataset (Subramanian et al., 2024, 2023; Devika et al., 2024) of news in Malayalam, with labels "half true", "partly false", "mostly false", and "false." The dataset consisted of social media posts in pure and code-mixed Malayalam, both in English and Malayalam scripts.

The details of the dataset are given in Table 3.

Label	Train set	Test set
FALSE	1386	100
MOSTLY FALSE	295	56
HALF TRUE	162	37
PARTLY FALSE	57	7

Table 3: Dataset details for the multi-class classification sub-task

We see that the train and test datasets have significant data imbalance, with "partly false" entries being roughly an order of magnitude less frequent. This makes classification more challenging, given the lack of cases that can teach the model a clear distinction between minority and other classes. We tried the same models as the binary classification sub-task with the same hyperparameters. The only exception was our prompt for GPT-4o, which was different from the binary classification sub-task. The GPT-4o is described below:

(system) You are an NLP expert helping classify fake news in Kerala. Before outputting, you will think what the text means within the cultural context of a Malayali. The categories like false, half true, etc. will tell how trustworthy the news text is. For example, 'half true' means the text is half true. Follow reasoning like this:

1. Think about what this sentence means, and put in the larger societal context of Kerala.

2. Revisit the training examples, and check whether your prediction agrees with the kind of labels that the training examples have.

3. Make sure you choose your final answer after carefully weighing the possibilities, for example, is it 'mostly false' or 'false'.

(user) You are a classifier. Use the training data below to classify each text as [FALSE, MOSTLY FALSE, HALF TRUE, PARTLY FALSE], output only a json that is a list of records with fields 'text' and 'prediction':

Table 4 summarizes the performance of our models we tested for this sub-task. We see that GPT-4o performed the best, and TFIDF + Logistic Regression did better than Fasttext and Malayalam BERT, a surprising result. However, the train-test performance gap is higher for logistic regression. This needs further exploration, as a better choice of hyperparameters combined with synthetic data or undersampling the majority classes might help improve the test set macro F1. We experimented with synthetic data generated by GPT-4o, but it did not yield noticeable performance improvements.

Model	Train Macro F1	Test Macro F1
TFIDF + Log. Reg.	0.297	0.203
FastText	0.209	0.167
`malayalam-bert`	0.209	0.167
GPT-4o	-	**0.290**

Table 4: Train and Test set Macro F1 scores for the multi-class classification sub-task

4 Detecting AI-generated product reviews

4.1 Dataset and Task Description

The dataset provided by the organizers for sub-task 5 (Premjith et al., 2025) was divided into separate subsets for Tamil and Malayalam. Each dataset comprised a mix of human-generated and machine-generated product reviews. The objective of this task is to develop and evaluate models capable of accurately distinguishing between AI-generated and human-generated reviews in these languages, effectively addressing a binary classification problem. The distribution of classes is shown below.

Language	Human Gen.	AI Gen.
Malayalam	400	400
Tamil	403	405

Table 5: Class Distribution for Tamil and Malayalam Datasets

4.2 Methods

We adopted a fine-tuning approach using several transformer-based encoder and decoder models. Our base models included the multilingual BERT base model (Devlin et al., 2019b), two monolingual BERT models released by L3Cube (Joshi, 2022), and GPT-2 (Radford et al., 2019). The multilingual BERT model was pre-trained on 102 languages with masked language modeling (MLM) and next sentence prediction (NSP) objectives. The monolingual BERT models were fine-tuned from the existing multilingual model using a monolingual corpus. For GPT-2, we utilized the 124M parameter version model, a transformer-based decoder-only model pre-trained on a large dataset with a causal language modeling (CLM) objective.

The fine-tuning process involves initializing each model with pre-trained weights and adding a classification head on top with 10% dropout. For each subtask, we fine-tune the entire model on the given dataset using stochastic gradient descent with back-propagation. As the subtask is a binary classification problem, we use cross-entropy loss as the objective function. Each dataset is split into 80% for training and 20% for testing, and we utilize the ADAM optimizer (Kingma and Ba, 2017) with an exponential learning rate scheduler. Table 6 shows the training hyperparameters we used.

Parameter	**Value**
Learning rate	5e-5
Learning rate decay	0.9
Batch size	32
Training epochs	10

Table 6: Training hyperparameters

4.3 Results

We fine-tuned all three models on the training set for 10 epochs. During training, we tracked accuracy, precision, F1 scores, and training loss. In our experiment `bert-base-multilingual`, `tamil-bert`, and `malayalam-bert` outperformed GPT-2 in all their respective tasks.

Tables 7 and 8 show the observed results for the Tamil and Malayalam language task respectively.

Tables 9 shows the performance of our submitted models and overall ranks on the held-out test set. The difference in macro F1 scores between our test set and the held-out test set suggests that the final test set included more complex and subtle texts, which posed greater challenges for our fine-tuned model to identify effectively.

Metric	bert-base-multi	tamil-bert	gpt2
Accuracy	0.9938	0.9877	0.9568
Precision	0.9885	0.9773	0.9438
Recall	1.0000	1.0000	0.9767
F1-Score	**0.9942**	0.9885	0.9600

Table 7: Evaluation of fine-tuned models on Tamil test set

Metric	bert-base-multi	malayalam-bert	gpt2
Accuracy	0.9688	0.9688	0.9062
Precision	0.9870	0.9630	0.9012
Recall	0.9500	0.9750	0.9125
F1-Score	0.9682	**0.9689**	0.9068

Table 8: Evaluation of fine-tuned models on Malayalam test set

Language	Macro F1	Rank
Malayalam	0.9	4/51
Tamil	0.646	30/54

Table 9: Evaluation of fine-tuned models on held-out test set

5 Conclusion

In this paper we explore modeling approaches for detecting fake news in Malayalam, a low-resource Dravidian language, and tackle the challenge of identifying AI-generated product reviews in Malayalam and Tamil.

For detecting fake news in Malayalam, we explored several approaches within the binary and multi-class classification sub-tasks. We discovered that while simpler approaches like Fasttext yield good performance on the test set, the overfitting is much higher than fine-tuned transformer models such as `malayalam-bert`. In multi-class classification, we were unable to achieve significant macro F1s and saw that GPT-4o did the best, indicating the need for further exploration and error analysis.

For detecting AI-generated product reviews we tested several transformer-based models, including multilingual and monolingual BERT models and GPT-2, fine-tuning them on provided datasets. Our results showed that BERT-based models outperformed GPT-2 in most cases.

6 Limitations

While our paper compares various models and discusses their ability in detecting fake news and AI-generated product reviews in Dravidian languages, there are certain limitations as well. Fake news on social media and AI-generated content in e-commerce are rapidly evolving issues, which are more difficult to detect in lower-resource languages and for communities with complex cultural nuances and rapidly evolving social landscapes. These challenges are further compounded by the continuous changes in language use and the emergence of new forms of disinformation, making it essential for models to be adaptive to shifting patterns. Approaches such as active and continual learning and other qualitative feedback mechanisms are necessary to combat such issues, as they enable the system to update its knowledge base and improve its accuracy over time. Furthermore, real-world challenges such as domain adaptation present additional difficulties; models trained on one domain may not perform optimally when transferred to another, due to differences in context, vocabulary, and cultural references. Moreover, such NLP-based systems might be biased towards or against certain views, thus unintentionally suppressing the opinions of well-meaning individuals. These biases can emerge due to the data used to train models, which may not be representative of diverse viewpoints or communities. The lack of diverse training data can inadvertently lead to the marginalization of certain demographic groups. Larger unsupervised and supervised datasets are necessary to capture such nuances, in order to avoid socioeconomic biases in online platforms using models described in this paper. Additionally, the presence of these biases can affect the real-world applicability of these models, as they may produce skewed results when deployed in different contexts or for different populations. Furthermore, given the black-box nature of the models in our paper, we also need to focus on investigating their interpretability and explainability. Understanding how these models arrive at their decisions is crucial for addressing potential biases and improving their fairness, as well as for fostering trust among users and stakeholders.

References

Sankalp Bahad, Yash Bhaskar, and Parameswari Krishnamurthy. 2024. Fine-tuning language models for AI vs human generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 918–921, Mexico City, Mexico. Association for Computational Linguistics.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. Indicbart: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics.

K Devika, B Haripriya, E Vigneshwar, B Premjith, Bharathi Raja Chakravarthi, et al. 2024. From dataset to detection: A comprehensive approach to combating malayalam fake news. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 16–23.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL 2019*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Arpita Dhar and Anjali Agarwal. 2018. Fake news detection using deep learning models. In *Proceedings of the 2018 ICML*, volume 97, pages 1008–1018. PMLR.

Arpita Ghosh and Pabitra Mitra. 2017. Detecting fake news in social media: A data mining perspective. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 743–752. ACM.

Farrokh Habibzadeh. 2023. GPTZero performance in identifying artificial intelligence-generated medical texts: A preliminary study. *J. Korean Med. Sci.*, 38(38):e319.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2024. A watermark for large language models. *Preprint*, arXiv:2301.10226.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019.

Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Jiwei Luo, Guofang Nan, Dahui Li, and Yong Tan. 2023. Ai-generated review detection. *Available at SSRN 4610727*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *Preprint*, arXiv:2301.11305.

B Premjith, Nandhini K, Bharathi Raja Chakravarthi, Thenmozhi Durairaj, Balasubramanian Palani, Kumaresan Thavareesan, Sajeetha, and Prasanna Kumar. 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, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

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.

Kai Shu, Anna Sliva, Siyi Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 1–10. SIAM.

Malliga Subramanian, , B Premjith, Kogilavani Shanmugavadivel, Santhia Pandiyan, Balasubramanian Palani, and Bharathi Raja Chakravarthi. 2025. Overview of the Shared Task on Fake News Detection 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.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, B Premjith, K Vanaja, S Mithunja, K Devika, et al. 2024. Overview of the second shared task on fake news detection in dravidian languages: Dravidianlangtech@ eacl 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 71–78.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, Muskaan Singh, Sandhiya Raja, Vanaja, and Mithunajha S. 2023. Overview of the shared task on fake news detection from social media text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.