Girma@DravidianLangTech 2025: Detecting AI Generated Product Reviews

Girma Yohannis Bade[1,a], **Muhammad Tayyab Zamir**[2,a], **Olga Kolesnikova**, **José Luis Oropeza**[4,a], **Grigori Sidorov**[5,a], **Alexander Gelbukh**[6,a]
[a]Centro de Investigaciones en Computación(CIC),
Instituto Politécnico Nacional(IPN), Miguel Othon de Mendizabal,
Ciudad de México, 07320, México.
[1]girme2005@gmail.com

Abstract

The increasing prevalence of AI-generated content, including fake product reviews, poses significant challenges in maintaining authenticity and trust in e-commerce systems. While much work has focused on detecting such reviews in high-resource languages, limited attention has been given to low-resource languages like Malayalam and Tamil. This study aims to address this gap by developing a robust framework to identify AI-generated product reviews in these languages. We explore a BERT-based approach for this task. Our methodology involves fine-tuning a BERT-based model specifically on Malayalam and Tamil datasets. The experiments are conducted using labeled datasets that contain a mix of human-written and AI-generated reviews. Performance is evaluated using the macro F1 score. The results show that the BERT-based model achieved a macro F1 score of 0.6394 for Tamil and 0.8849 for Malayalam. Preliminary results indicate that the BERT-based model performs significantly better for Malayalam than for Tamil in terms of the average Macro F1 score, leveraging its ability to capture the complex linguistic features of these languages. Finally, we open the source code of the implementation in the GitHub repository: AI-Generated-Product-Review-Code.

Keywords: AI-generaated, Detection, Product review, BERT

1 Introduction

The e-commerce marketplaces has led to an increase in online product reviews, which have become important for consumer buying behavior. Nonetheless, the problem of AI content generation has worsened the trust issues surrounding platforms due to the flood of fake reviews. These types of review can deceive users, affect the company's image, and alter competitive structures, making it difficult to devise new methods to identify them (Ott et al., 2011; Banerjee and Chua, 2014). With so many efforts being made to identify fake reviews in the English language, not many focus is acquired towards low-resource languages such as Malayalam and Tamil. This gap needs to be filled as these low resource languages are unique in their own way.

NLP does not tend to focus much on Malayalam and Tamil, which are common in South India as well as with the diaspora because resources are limited (Joshi et al., 2020; Zamir et al., 2024a).

The goal for this particular research is to create an effective hybrid system for the detection and classification of AI-produced reviews in Malayalam and Tamil using modern transformer. In particular, we utilize a BERT-based model which has been fine tuned to Malayalam and Tamil datasets. To explain it in simple terms, BERT is a great performer in most NLP learning tasks because it uses context and other deep language features to understand text's meaning (Devlin, 2018; Ahani et al., 2024), and we tend to fine tune BERT to specific low resource languages datasets to help it perform even better.

The described work requires the collection, and careful labeling of approximately authentic-sounding reviews of AI technology in both languages, which have been written in a mix of human and AI text. Both training and evaluation of the models are performed using the Macro F1-score, and their changes. Initial analysis indicates that the BERT-based model ultimately achieves the best results for Malayalam language as compared to Tamil language baseline making it a great alternative for cases involving complex linguistic phenomena (Ullah et al., 2024).

2 Related Works

With the advanced evolution of AI-generated text models, such as GPT-4 and its future models, the content generation process has changed in several spheres, including reviews of online products. Al-

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 228–233

though there's an enormous opportunity in these tools, their inappropriate use has brought forth concerns regarding authenticity, especially in sensitive fields like consumer decision-making. The problem becomes more pronounced in low-resource languages like Malayalam and Tamil since there is still very little research done in an AI generated content detection. Other literature point towards resource rich languages, like English, and emphasize on the ethical consideration as well as the systems in place of text generation detection systems (Solaiman et al., 2019). Nevertheless, the application of these techniques in low resource languages poses a problem owing to the distinctive linguistic features of these languages.

Malayalam and Tamil belong to low resource languages that is characterized by its high morphology, agglutinative nature and sophisticated syntactic features. Research has demonstrated that such languages are a very difficult case for natural language processing (NLP) owing to suffixation, highly compound words and poor availability of standard transliteration (Annamalai, 2010; Chakravarthi et al., 2022b) which makes it hard to use many methods designed for resource-rich languages without major adaptations.

This is why such models cannot simply be transferred directly from resource-rich languages as they require specialized approaches to help, for example text classification or AI related content detection.

While datasets are the be-all and end-all for training detection models, Dravidian languages have short supply of annotated data. One example here is the "Dravidian CodeMix" shared task (Chakravarthi et al., 2023; Tash et al., 2024) which provides a dataset of code-mixed sentiment analysis and offensive language detection. Especially, there are few datasets being curated for AI-generated content detection in Dravidian languages. Thus, works such as "Shared Task for Detector AI-generated Product Reviews in Dravidian Languages" fill this need by providing training and test datasets in Malayalam as well Tamil, allowing researches to train their models suited for these languages.

Methods (when it comes to detecting AI-generated text) were primarily based on linguistic characteristics analysis and stylometry like n-gram comparation, detecting a style inconsistence (Jawahar et al., 2019). Recent transformer powered models like BERT, RoBERTa and mBERT have established excellent performance in text classification tasks due to the impact of deep learning. Generalizable Multilingual models (XLM-R (Jawahar et al., 2019; Chakravarthi et al., 2022a)) that outperforms in low-resource languages, are showing an encouraging performance when they are finetuned on domain-specific data. Performance of AI-based text detection models is evaluated based on evaluation metrics (especially when we only care about identifying the text generated by AI), and developing such models is no exception.

In NLP, the score of F1 is an accuracy metric that strikes a balance between recall and precision (Naidu et al., 2023), reflecting on the other hand it was suited for classification problems with imbalanced datasets (Bade et al., 2024c). Adoption of the metric in this Shared Task underscores the need for a more comprehensive model performance evaluation measure for low-resource languages (Priyadharshini et al., 2022; Zamir et al., 2024b). One of the biggest ethical concerns in identifying AI-generated content detection is within low resource languages, to keep trust on online platforms alive. Kimera et al. (2024) has claimed that systematic detection systems can help to prevent false information and building trustworthy digital eco-system. In future work, we will address issues of more diverse and representative datasets (including multimodal), as well as breaking biases to improve detection systems for these languages.

3 System Description

In this section, we discuss about datasets, preprocessing, feature extraction, and model seletion. Moreover, finally it overviews architecture of this task.

3.1 Datasets

The research in the NLP domain heavily relies on well-curated datasets, which serve as the driving force for creating intelligent systems (Bade et al., 2024a). However, it is labor intensive to obtain well-written data to train the language model, especially under-resourced ones (Bade, 2021). Thanks to DravidianLangTech (Priyadharshini et al., 2023), they offered datasets and task instructions for this work (Premjith et al., 2025). The datasets are organized into two subsets: training and test sets. The training data set is a data set that contains two variables, X input and Y output. While the X variable represents users' comments, the Y variable repre-

sents their values that can determine whether the comments are human written or machine generated. However, the test dataset does not contain the Y variable because we expect it to be predicted by the model. This kind of dataset arrangements are more conveinent for supervised machine learning. Mathematically,

$$\text{Training Data} = \sum_{j=1}^{n} (X_{ij}, Y_{ij})$$

and

$$\text{Test Data} = \sum_{j=1}^{m} X_{ij}$$

Table 1 presents the detailed statistics of these datasets.

Languages	Dataset	has_label?	Size
Tamil	Train	yes	808
	Test	no	100
	Total	**–**	**908**
Malayalam	Train	yes	800
	Test	no	200
	Total	**–**	**1,000**

Table 1: Dataset statistics

Table 1 outlines the training and test datasets. While training dataset served as the primary resource for training the selected algorithm, test dataset served to evaluate the final performance of the model. Notably, test data was kept separate and remains unseen during the training process. Table 2 further provides the class label distribution for the training.

Language	Label	Count
Tamil	HUMAN	403
	AI	405
	Total	**808**
Malayalam	HUMAN	400
	AI	400
	Total	**800**

Table 2: The statistics of class label distributions of the training dataset.

3.2 Preprocessing

The annotated training, development datasets, and test dataset underwent pre-processing. Then punctuation mark removal, emoji removal, and username removal are the main objectives of this step in this particular use case. The built-in "re" module in Python has helped to eliminate all these staff.

3.3 Feature Extraction

Since AI algorithms operate on numeric data (Bade and Seid, 2018), it is necessary to encode the input of text to a numeric equivalent (Bade et al., 2024d). The process of converting text input into numeric form is known as data encoding or feature extraction (Bade et al., 2024a). This task is carried out by BertTokenizer of the BERT model.

3.4 Model Selection

Once the NLP processing steps such as dataset organization, pre-processing, and feature extraction are completed, the next critical step involves selecting and applying appropriate AI algorithms. In this study, we employ the BERT model, a state-of-the-art architecture based on Transformers, to achieve our objectives (Yigezu et al., 2023), specifically the bert-base-uncased variant. Since its introduction, Transformers have revolutionized NLP by enabling enhanced parallelization and effectively capturing long-range dependencies (Vaswani, 2017). Among these models, BERT remains a fundamental baseline, consistently achieving state-of-the-art performance across various NLP benchmarks (Rogers et al., 2021).

To process textual data, BERT employs its dedicated BertTokenizer, which tokenizes text and converts it into numerical representations, ensuring efficient input processing for the model (Bade et al., 2024b). Table 3 outlines the hyperparameters used for this model.

Hyperparameters	Values
Learning Rate	1e-5
Evaluation Strategy	Epoch
Epochs	5
Batch Size	32
Activation function@output level	Sigmoid

Table 3: BERT Hyperparameters

As we can see from Table 3, the learning rate indicates the number of times the exectuion taken place to improve the model's performance. Epoch refers to one complete pass through the entire training dataset by the learning algorithm (Mersha et al., 2024). Thus, we set the epoch to be 5,.i.e the execution did pass 5 complete times. The batch size refers to dividing the total data size into 32 and

bringing the divided batch one a time for the execution. This helps the execution to be fast. The last parameter, activation function is used to label or group the computational result into two classes. Figure 1 presents the workflow of this study.

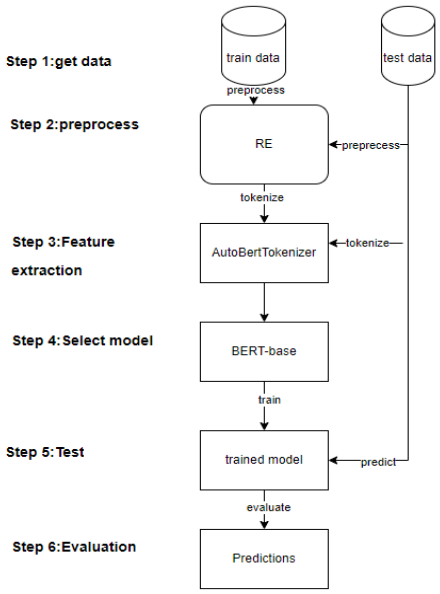

Figure 1: Overall workflow of the study.

In Figure 1, the pipeline begins with acquiring a training dataset, followed by pre-processing and feature extraction steps. The selected model, BERT-base, is then trained to learn the patterns and behavior specific to this dataset. Once training is complete, the model is evaluated using test data to generate predictions.

4 Result and Discussion

We trained the provided dataset on the BERT model and tested its performance using a separate test set. The predictions generated from the test set data were submitted to the workshop organizers for evaluation. These submissions were evaluated using accuracy (Acc), average macro precision (P), average macro recall (R), and average macro F1-score (F1) as evaluation metric. The final results, published by the organizers, revealed that for Tamil, BERT attained a significantly higher macro F1 score of 0.6394, and for Malayalam, it excelled with a macro F1 score of 0.8849. Table 4 shows more details.

From Table 4, we can easily infer that the selected model and configured hyperparameters are more favored for Malayalam than Tamil.

Language	Acc	P	R	F1
Tamil	0.6400	0.6394	0.6394	0.6394
Malayalam	0.8850	0.8859	0.8850	0.8849

Table 4: Performance metrics of **BERT** model for the data of Malayalam and Tamil languages.

5 Comparative Analysis

The suggested AI-generated review detection model is contrasted with other baseline techniques, such as naive Bayes, support vector machines (SVMs), BERT, ALBERT, RoBERTa, and gradient boosting decision trees (GBDTs).These baseline methods are compared to our approaches in terms of the F1 measure, regardless of the dataset they employed.

Model	F1-score
Qualitative (Fröhnel et al., 2025)	0.5300
GANs(Ke et al., 2025)	0.9500
RoBERTa (Wang et al., 2025)	0.7342
BERT (**Ours**)Mal	**0.8849**
BERT (**Ours**)Tam	0.6394

Table 5: Comparison of the models of our work and others. The result in bold shows the performance achieved by our approach, revealing the effectiveness of the model.

6 Conclusion and Future Work

This research created a methodology for detecting reviews generated by AI for products in Malayalam and Tamil using a fine-tuned BERT model. The experimental results proved that the BERT model is the best performer and scored 0.8849 on Macro F1-score for the Malayalam language compared to 0.6394 scored by the Tamil language. The findings support the claim regarding the sophisticated grammatical features possessed by BERT for these low resource languages, which makes the translation of these languages into other languages rather appealing due to the challenges posed by the insufficient resources. This work highlights the fact that the models for advanced languages are necessary for the impoverished context and attempts to provide tools that could be used to improve the language's authenticity.

In the future, this work will be improved both by adding more diverse domains and by using multilingual models in potential cross-lingual transfer learning setups. BERT can also be enhanced by

using explainable AI techniques. This will help to increase the accuracy of detection and support broader use cases in combating AI-generated content.

Limitation and Ethics Statement

The datasets for Tamil and Malayalam languages used in this study were limited in size, and the model was trained on this relatively small dataset. As a result, the observed performance may not generalize well to all unseen data. Despite these constraints, our model demonstrated comparable performance in detecting AI-generated product reviews for social media posts. Nevertheless, in a highly competitive environment, our method achieved impressive rankings of 10th and 32nd for Malayalam and Tamil, respectively. Additionally, our work adheres to the ethical principles outlined for computational research and professional conduct[1].

Acknowledgment

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

Z Ahani, M Tash, M Zamir, and I Gelbukh. 2024. Zavira@ dravidianlangtech 2024: Telugu hate speech detection using lstm. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 107–112.

Elay Annamalai. 2010. Politics of language in india. In *Routledge Handbook of South Asian Politics*, pages 213–231. Routledge.

Girma Bade, Olga Kolesnikova, Grigori Sidorov, and José Oropeza. 2024a. Social media hate and offensive speech detection using machine learning method. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 240–244.

Girma Yohannis Bade. 2021. Natural language processing and its challenges on omotic language group of ethiopia. *Journal of Computer Science Research*, 3(4):26–30.

Girma Yohannis Bade, O Koleniskova, José Luis Oropeza, Grigori Sidorov, and Kidist Feleke Bergene. 2024b. Hope speech in social media texts using transformer. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEURWS. org*.

Girma Yohannis Bade, Olga Kolesnikova, and Jose Luis Oropeza. 2024c. Evaluating the quality of data: Case of sarcasm dataset.

Girma Yohannis Bade, Olga Kolesnikova, José Luis Oropeza, and Grigori Sidorov. 2024d. Lexicon-based language relatedness analysis. *Procedia Computer Science*, 244:268–277.

Girma Yohannis Bade and Hussien Seid. 2018. Development of longest-match based stemmer for texts of wolaita language. *vol*, 4:79–83.

Snehasish Banerjee and Alton YK Chua. 2014. Applauses in hotel reviews: Genuine or deceptive? In *2014 Science and Information Conference*, pages 938–942. IEEE.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Shubanker Banerjee, Manoj Balaji Jagadeeshan, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sean Benhur, and John Philip McCrae. 2023. Detecting abusive comments at a fine-grained level in a low-resource language. *Natural Language Processing Journal*, 3:100006.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Subalalitha Cn, Sangeetha S, Malliga Subramanian, Kogilavani Shanmugavadivel, Parameswari Krishnamurthy, Adeep Hande, Siddhanth U Hegde, Roshan Nayak, and Swetha Valli. 2022a. Findings of the shared task on multi-task learning in Dravidian languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 286–291, Dublin, Ireland. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 2022b. Overview of the shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 369–377, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[1] https://www.aclweb.org/portal/content/acl-code-ethics

Kim Fröhnel, Bennet Santelmann, and Rüdiger Zarnekow. 2025. Genuine or fake? explaining consumers' perception and detection of ai-generated fake reviews.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Zong Ke, Shicheng Zhou, Yining Zhou, Chia Hong Chang, and Rong Zhang. 2025. Detection of ai deepfake and fraud in online payments using gan-based models. *arXiv preprint arXiv:2501.07033*.

Richard Kimera, Yun-Seon Kim, and Heeyoul Choi. 2024. Advancing ai with integrity: Ethical challenges and solutions in neural machine translation. *arXiv preprint arXiv:2404.01070*.

Melkamu Abay Mersha, Girma Yohannis Bade, Jugal Kalita, Olga Kolesnikova, Alexander Gelbukh, et al. 2024. Ethio-fake: Cutting-edge approaches to combat fake news in under-resourced languages using explainable ai. *Procedia Computer Science*, 244:133–142.

Gireen Naidu, Tranos Zuva, and Elias Mmbongeni Sibanda. 2023. A review of evaluation metrics in machine learning algorithms. In *Computer Science On-line Conference*, pages 15–25. Springer.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*.

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.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and Prasanna Kumar Kumaresan. 2023. Findings of the shared task on Abusive Comment Detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages DravidianLangTech 2023*. Recent Advances in Natural Language Processing.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

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

M Tash, Z Ahani, M Zamir, O Kolesnikova, and G Sidorov. 2024. Lidoma@ lt-edi 2024: Tamil hate speech detection in migration discourse. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 184–189.

Fida Ullah, Muhammad Zamir, Muhammad Arif, M Ahmad, E Felipe-Riveron, and Alexander Gelbukh. 2024. Fida@ dravidianlangtech 2024: A novel approach to hate speech detection using distilbert-base-multilingual-cased. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 85–90.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Yuxia Wang, Artem Shelmanov, Jonibek Mansurov, Akim Tsvigun, Vladislav Mikhailov, Rui Xing, Zhuohan Xie, Jiahui Geng, Giovanni Puccetti, Ekaterina Artemova, et al. 2025. Genai content detection task 1: English and multilingual machine-generated text detection: Ai vs. human. *arXiv preprint arXiv:2501.11012*.

Mesay Gemeda Yigezu, Girma Yohannis Bade, Atnafu Lambebo Tonja, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023. Bilingual word-level language identification for omotic languages. In *International Conference on Advances of Science and Technology*, pages 63–77. Springer.

M Zamir, M Tash, Z Ahani, A Gelbukh, and G Sidorov. 2024a. Tayyab@ dravidianlangtech 2024: detecting fake news in malayalam lstm approach and challenges. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 113–118.

Muhammad Tayyab Zamir, Fida Ullah, Rasikh Tariq, Waqas Haider Bangyal, Muhammad Arif, and Alexander Gelbukh. 2024b. Machine and deep learning algorithms for sentiment analysis during covid-19: A vision to create fake news resistant society. *PloS one*, 19(12):e0315407.