NLP_goats@DravidianLangTech 2025: Detecting Fake News in Dravidian Languages: A Text Classification Approach

Srihari V K
Sri Sivasubramaniya Nadar College of Engineering
srihari2210434@ssn.edu.in

Vijay Karthick Vaidyanathan
Sri Sivasubramaniya Nadar College of Engineering
vijaykarthick2210930@ssn.edu.in

Thenmozhi Durairaj
Sri Sivasubramaniya Nadar College of Engineering
theni_d@ssn.edu.in

Abstract

The advent and expansion of social media have transformed global communication. Despite its numerous advantages, it has also created an avenue for the rapid spread of fake news, which can impact people's decision-making and judgment. This study explores detecting fake news as part of the DravidianLangTech@NAACL 2025 shared task, focusing on two key tasks. The aim of Task 1 is to classify Malayalam social media posts as either original or fake, and Task 2 categorizes Malayalam-language news articles into four levels of truthfulness: False, Half True, Mostly False and Partly False. We accomplished the tasks using transformer models, e.g., mBERT and classifiers like Naive Bayes. Our results were promising, with mBERT achieving the better results. We achieved a macro-F1 score of 0.83 for distinguishing between fake and original content in Task 1 and a score of 0.54 for classifying news articles in Task 2, ranking us 11 and 4, respectively.

1 Introduction

The rapid growth of social networks has transformed communication, allowing users to express opinions and share content easily. However, this has led to the spread of fake news, defined as false or misleading information presented as real news (Shu et al., 2017). The viral nature of social media amplifies its spread, influencing decisions and eroding trust in genuine sources (Vosoughi et al., 2018), highlighting the need for effective detection mechanisms (Kumar and Shah, 2018).

This challenge is even greater for underrepresented languages such as Dravidian languages, including Malayalam, Tamil, and Telugu, which face a lack of computational resources and annotated datasets for NLP. To address this, DravidianLangTech@NAACL 2025 proposes two tasks: Task 1 classifies the text of social media as fake or original, and Task 2 detects fake news in Malayalam News. These tasks aim to improve fake news detection in Dravidian languages using large datasets and advanced machine learning techniques.

Section 2 reviews related work on fake news detection, especially for Dravidian languages. Section 3 describes the task, and Section 4 outlines the methodology, including details of the data set and the model selection. Section 5 presents experimental results, and Section 6 offers error analysis. Section 7 concludes by summarizing the study's findings.

This study aims to enhance misinformation detection in low-resource languages, focusing on Dravidian languages, by developing effective methods for identifying fake news using models like mBERT and Naive Bayes.

The taken approach is not specific to only detecting fake news, but can also be used to detect abusive language in text towards women as shown in (Rajiakodi et al., 2025). The goal is to contribute to the advancement of NLP and misinformation detection. For implementation, please refer to this GitHub repository (srihari2704).

2 Related Work

The detection of fake news continues today. An evolving field, it has ample scope for improvement, especially in languages with fewer resources and datasets. The improvement in technology and extensive research combined with the availability of datasets for such languages have improved the performance of models to detect fake news in such languages.

The author (Shu et al., 2017) examined methods that integrate content-based characteristics and social context characteristics, highlighting the importance of modeling user behavior and social network analysis. This study opened the door to integrating contextualized information alongside text mining and motivated selected models that are multimodal and/or hybrid.

A pivotal study, (Wang, 2017), introduced the LIAR dataset. A widely used benchmark for fake news detection, this work used logistic regression and SVM classifiers. It showcased the value of curated datasets. It complements the work by Shu et al., providing the necessary data infrastructure to evaluate approaches combining content and social context.

Using deep learning approaches, (Abualigah et al., 2024) presented the power of neural networks in extracting semantic features for text content. This paper showed how deep learning architectures, particularly Bidirectional LSTMs (BiLSTMs), could generalize better than classical classifiers if combined with the linguistically rich word embeddings GloVe. Building on the earlier focus on features and datasets, this study transitioned the field toward neural approaches for feature extraction and classification.

A multifaceted strategy to counteract Malayalam fake news, (Devika et al., 2024a) extended the work on data set curation by introducing a labeled dataset specifically for Malayalam fake news. Their adoption of multilingual BERT and machine learning classifiers highlighted the challenges of generalizing state-of-the-art techniques to resource-poor environments. This is consistent with using datasets and pre-trained models for language-specific tasks.

The author (Rahman et al., 2024) further emphasized the power of language-specific models, as they obtained the best-shared task F1 score. These studies prove that advanced neural architectures can excel when adapted appropriately.

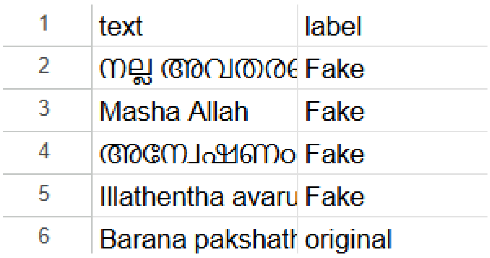
These studies provide a clear overview of the trajectory of fake news detection. Early research revolved around datasets, feature design, and simple classifiers; further work incorporated deep learning techniques and language-specific approaches. The shift toward such low-resource languages as Malayalam marks a move toward linguistic diversity and tailoring advanced technology to address a global issue.

3 Task Description

The shared task of Fake News Detection in Dravidian Languages aims to address the widespread misinformation on online platforms using given datasets. It is divided into two subtasks:

3.1 Task 1

Classify English social media posts or comments from Twitter, Facebook, and YouTube as *fake* or *original*. The dataset as shown in Figure 1 for this task is provided by previous work on the detection of fake news in Dravidian languages (Subramanian et al., 2025, 2023). Participants develop systems to identify misinformation and ensure the authenticity of online content.

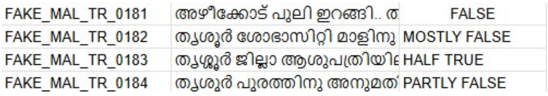


1	text	label
2	നല്ല അവതര	Fake
3	Masha Allah	Fake
4	അന്വേഷണം	Fake
5	Illathentha avaru	Fake
6	Barana pakshatl	original

Figure 1: Dataset for task 1

3.2 Task 2

Detect and categorize Malayalam news articles into four labels: *False, Half True, Mostly False, Partly False.* The task focuses on addressing misinformation in regional languages to ensure inclusivity and accuracy in local news verification. The dataset as shown in Figure 2 for this task is based on previous studies (Subramanian et al., 2024; Devika et al., 2024b).



FAKE_MAL_TR_0181	അഴീക്കോട് പുലി ഇറങ്ങി.. ത	FALSE
FAKE_MAL_TR_0182	തൃശൂർ ശോഭാസിറ്റി മാളിനു	MOSTLY FALSE
FAKE_MAL_TR_0183	തൃശ്ശൂർ ജില്ലാ ആശുപത്രിയി	HALF TRUE
FAKE_MAL_TR_0184	തൃശൂർ പൂരത്തിനു അനുമതി	PARTLY FALSE

Figure 2: Dataset for task 2

4 Methodology

This study designs machine learning models to classify social media comments in the Malayalam language as fake or real, thus countering misinformation. The dataset consisted of labelled comments for supervised learning. Pre-trained multilingual BERT is used to tokenize and process the text for model training. The dataset is split into training and validation sets, and the model is fine-tuned to detect fake news. These performances are reviewed based on accuracy, precision, recall, and F1 score. The goal is to reduce the spread of misinformation and encourage informed discussion in digital communication.

4.1 Data Preprocessing

Preprocessing played a critical role in dataset preparation, enabling better results when predicting fake news. The train and development datasets, each with two columns (text and label), were loaded using Pandas. A quick inspection verified their structure and integrity.

Tables 1 and 2 show that the label distribution was analyzed using Matplotlib bar charts to identify possible class imbalances. To address this, random oversampling was employed to balance the classes in Task 2, as it exhibited a significant imbalance. However, Task 1 was already relatively balanced, so no oversampling was performed. Labels were encoded into binary format, with "Fake" zero and "Original" 1, to ensure consistency and compatibility with machine learning models.

Label	Count
Original	1658
Fake	1599

Table 1: Label distribution of Task 1 dataset.

Label	Count
FALSE	1386
MOSTLY FALSE	295
HALF TRUE	162
PARTLY FALSE	57

Table 2: Label distribution of Task 2 dataset.

Text cleaning removed noise, including punctuation and emojis, using regular expressions to enhance the clarity of the data set. Following cleaning, the text was reviewed to confirm accurate preprocessing.

HuggingFace's AutoTokenizer from the multilingual BERT model (bert-base-multilingual-cased) was used for tokenization, incorporating padding and truncation to 128 tokens for computational efficiency. The datasets were converted into HuggingFace's Dataset format, enabling seamless integration with the training pipeline.

4.2 Model Description

Fake news detection is tackled using various machine learning algorithms and transformer models. Traditional classifiers offer baseline comparisons, while mBERT enhances performance with contextual understanding, allowing evaluation of multiple approaches to identify the most effective solution.

4.2.1 Model selection

Table 3 compares different machine learning and deep learning models for detecting fake news in Malayalam in terms of their performance scores. Among them, M-BERT has the best score (0.85), which shows that it is superior in detecting contextual subtleties in the Malayalam language. Naïve Bayes comes second with a score of 0.80, demonstrating that probabilistic techniques are still influential. Conventional machine learning models such as Logistic Regression, SVM, and MLP demonstrate competitive performance (between 0.77 and 0.79), underscoring their dependability in the face of the increasing popularity of deep learning methods. XLM-R, another transformer-based model, has a score of 0.76, marginally lower than M-BERT but still showing effectiveness. The findings show that transformer-based models such as M-BERT and XLM-R perform better than conventional approaches, supporting contextual embeddings' significance in addressing the Malayalam language's intricacies.

Model	Task1 Score	Task2 Score
Logistic Regression	0.79	0.79
Neural Network (MLP)	0.77	0.77
Support Vector Machines (SVM)	0.78	0.78
MLP (Alternate)	0.72	0.73
Naïve Bayes	0.80	0.80
XLM-R	0.76	0.77
M-BERT	0.85	0.85

Table 3: Performance Comparison of Models for Fake News Detection in Malayalam

4.2.2 mBERT

Multilingual BERT (mBERT) is an extension of BERT designed to manage multilingual inputs. Introduced in (Devlin et al., 2018), it is trained on a corpus that includes 104 languages, enabling cross-language predictions without needing separate models for each language (Devika et al., 2024a). Built on the transformer architecture, mBERT employs self-attention mechanisms to capture contextualized word embeddings across various languages, making it particularly effective for tasks such as fake news detection in Dravidian languages (Wu and Dredze, 2019).

In Task 1, which involves binary classification of social media posts as 'fake' or 'real,' mBERT performs well due to its contextual understanding and ability to generalize across languages.

Table 4 presents the performance report for the mBERT model in Task 1. The model achieved an accuracy of 0.85, demonstrating its effectiveness in distinguishing between fake and original news.

Metric	Value
Precision	0.8210
Recall	0.8973
F1 Score	0.8575
Accuracy	0.8503

Table 4: Performance Metrics of mBERT for Task 1

For Task 2, where fine-grained classification of Malayalam text is required (such as categorizing posts as "False" or "Half True"), mBERT demonstrates strong performance. Using its multilingual capabilities effectively, it can handle complex linguistic patterns (Pires et al., 2019). Its proficiency in processing low-resource languages highlights its robustness, particularly in data-scarce scenarios (Ruder et al., 2019).

Table 5 displays the mBERT model's performance report in Task 2. The model achieved an accuracy of .80, demonstrating its effectiveness in distinguishing between fake and original news.

Category	Precision	Recall	F1-Score
False	0.79	0.89	0.84
Half True	0.33	0.17	0.23
Mostly False	0.32	0.17	0.22
Partly False	0.08	0.10	0.09

Table 5: Performance Metrics of mBERT for Task 2

mBERT has limitations with underrepresented languages or with significant morphological variations. Fine-tuning for specific tasks, such as fake news detection in Malayalam, improves its performance, making it a valuable tool for multilingual NLP and scalable misinformation detection (Ruder et al., 2019).

5 Results

Among the models used for the detection of fake news, mBERT significantly outperformed the others. Naive Bayes achieved a slightly higher macro F1 Score of 80.23, while mBERT achieved a macro F1 Score of 85.12. This highlights mBERT's ability to effectively represent long-range linguistic structures and context-sensitive elements, particularly for Malayalam, where shallower models proved less effective. The results emphasize the advantages of using advanced transformer-based architectures like mBERT for tasks involving multilingual and morphologically rich languages.

6 Error Analysis

An analysis of the fake news detection task revealed that the Naive Bayes model achieved a macro F1-score of 80.23 but struggled with ambiguous language and overlapping features. In contrast, mBERT performed better, scoring 85.12, although it encountered challenges with rare linguistic constructs and code-mixed content. Both models had difficulty because of the morphological complexity of the Malayalam language and the informal nature of the social media texts. This emphasizes the need for domain-specific pretraining, improved fine-tuning, and exploration of hybrid approaches.

7 Limitations

The primary limitations of this study stem from the challenges associated with processing low-resource languages like Malayalam, which lack large-scale annotated datasets for training robust fake news detection models. Morphological complexity and code-mixed content further hinder model performance, as transformer models like mBERT may struggle with rare linguistic constructs. Additionally, the study relies heavily on pretrained multilingual models, which, while effective, may not fully capture the nuances of Malayalam compared to language-specific models. Another limitation is the difficulty in detecting nuanced misinformation categories, such as subtle satire or partially false claims, which require deeper semantic understanding. Lastly, data imbalance in some categories, particularly in multi-class classification, may have influenced model generalizability, necessitating more refined balancing techniques for improved performance.

8 Conclusion

The mBERT model was identified as the most effective for fake news detection, achieving a macro F1-score of 0.83 in binary classification and 0.54 in multi-class classification. Its ability to capture contextual and semantic nuances in Malayalam text enabled it to outperform Naive Bayes, which struggled with the language's complexities. This underscores the value of advanced language models and emphasizes the need for robust preprocessing and fine-tuning to combat misinformation effectively.

References

Laith Abualigah, Yazan Yehia Al-Ajlouni, Mohammad Sh. Daoud, Maryam Altalhi, and Hazem Migdady. 2024. Fake news detection using recurrent neural network based on bidirectional lstm and glove. *Social Network Analysis and Mining*, 14(1):40.

K Devika, B Haripriya, E Vigneshwar, B Premjith, Bharathi Raja Chakravarthi, et al. 2024a. 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.

K Devika, B Haripriya, E Vigneshwar, B Premjith, Bharathi Raja Chakravarthi, et al. 2024b. 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. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186. Association for Computational Linguistics.

Srijan Kumar and Neil Shah. 2018. False information on web and social media: A survey. *Proceedings of the 25th International Conference on World Wide Web Companion*, pages 553–558.

Rafael Pires, Eduardo N. Ribeiro, and Luis C. Lamb. 2019. How multilingual is multilingual bert? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*, pages 5006–5018. Association for Computational Linguistics.

Tanzim Rahman, Abu Raihan, Md. Rahman, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. CUET_DUO@DravidianLangTech EACL2024: Fake news classification using Malayalam-BERT. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 223–228, St. Julian's, Malta. Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Sebastian Ruder, Matthew E. Peters, Iryna Gurevych, and Alexander Kementchedjhieva. 2019. A survey of cross-lingual embeddings. *Journal of Artificial Intelligence Research*, 65:405–443.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Explorations*, 19(1):22–36.

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.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

William Y. Wang. 2017. Liar, liar pants on fire: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, pages 422–426.

F. Wu and M. Dredze. 2019. Social media as a sensor of public opinion. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*, 2019:1001–1010.