Cognitext@DravidianLangTech2025: Fake News Classification in Malayalam Using mBERT and LSTM

Shriya Alladi[1], **Bharathi B**[2]
[1] Department of Information Technology
[2] Department of Computer science and Engineering
Sri Sivasubramania Nadar College of Engineering
shriya2310406@ssn.edu.in
bharathib@ssn.edu.in

Abstract

Fake news detection is a crucial task in combating misinformation, particularly in underrepresented languages such as Malayalam. This paper focuses on detecting fake news in Dravidian languages using two tasks: Social Media Text Classification and News Classification. We employ a fine-tuned multilingual BERT (mBERT) model for classifying a given social media text into original or fake and an LSTM-based architecture for accurately detecting and classifying fake news articles in the Malayalam language into different categories.

Extensive preprocessing techniques, such as tokenization and text cleaning, were used to ensure data quality. Our experiments achieved significant accuracy rates and F1-scores. The study's contributions include applying advanced machine learning techniques to the Malayalam language, addressing the lack of research on low-resource languages, and highlighting the challenges of fake news detection in multilingual and code-mixed environments.

1 Introduction

The digital age has amplified the spread of misinformation, posing severe societal and political challenges. While extensive research exists on fake news detection for global languages like English, regional and low-resource languages such as Malayalam remain underexplored. Malayalam, a Dravidian language spoken in southern India, presents unique challenges due to its script, morphology, and prevalence of code-mixed content on social media platforms.

Previous works have shown the efficacy of models such as BERT and LSTM for fake news detection, particularly in multilingual and sequential text processing tasks. However, their application to Dravidian languages remains limited.

This paper addresses this gap by proposing two fake news detection models: an mBERT model for social media text classification and an LSTM-based architecture for classifying news articles. Task 1 involves handling multilingual, code-mixed data with a focus on accurate social media classification, while Task 2 emphasizes sequential dependencies in Malayalam news articles. These contributions aim to enhance fake news detection for low-resource languages and address the societal need to combat misinformation effectively.

The rest of the paper is organized as follows: Section 2 analyzes the related works done in previous research, and Section 3 discusses the dataset. Section 4 contains a detailed discussion of the proposed models used in the current work. Section 5 explains the experimental results. Section 7 highlights the limitations of the study, while Section 7.1 outlines potential future research directions. In Section 6, we conclude the paper. Finally, the Acknowledgment section expresses gratitude to contributors and funding sources.

2 Related Works

Recent research in fake news detection has focused on applying various machine learning algorithms to identify misleading information in online content. (Yuslee and Abdullah, 2021) explore the use of Naive Bayes for fake news detection, highlighting its effectiveness in classifying news articles by leveraging natural language processing (NLP) techniques, including TF-IDF and Count Vectorizer. (Krishna and Adimoolam, 2022) compared the performance of Decision Tree algorithms and Support Vector Machines (SVM) in detecting fake news, emphasizing the reliability and novelty of the Decision Tree approach for fake news detection in social media. Similarly, (Mugdha et al., 2020) evaluate machine learning algorithms, including Naive Bayes, for detecting fake news in Bengali, focusing on feature extraction and classification performance in regional languages. (Ruchansky

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 439–443

et al., 2017) introduce CSI, a hybrid deep model for fake news detection, combining deep learning techniques with traditional methods to improve classification accuracy. Furthermore, (Devlin et al., 2019) present BERT, a deep bidirectional transformer model that has significantly advanced language understanding, and has shown impressive results in text classification tasks, including fake news detection. (Bahad et al., 2019) propose a fake news detection model based on Bi-directional LSTM-Recurrent Neural Network, demonstrating its superiority over other models like CNN, vanilla RNN, and unidirectional LSTM in terms of accuracy for detecting fake news. These studies collectively showcase a variety of machine learning approaches—from classical methods like Naive Bayes and Decision Trees to advanced models like BERT and Bi-directional LSTM—demonstrating their applicability in detecting fake news across different languages and platforms.

3 Dataset Description

he dataset is sourced from the Fake News Detection in Dravidian Languages provided by DravidianLangTech@NAACL 2025 (Subramanian et al., 2024)(Devika et al., 2024)(Subramanian et al., 2023)(Subramanian et al., 2025).

Task 1 consists of a dataset having 4,072 rows, split between the training and validation sets. The training set contains 3,257 articles, while the validation set consists of 815 articles. These articles are labeled as either fake or original, providing a foundation for model training and evaluation in a supervised setting.

Category	Rows
Train Set	3,257
Validation Set	815

Table 1: TASK 1 Data Summary

Task 2 consists of a dataset having a total of 3,120 rows, divided into a training set, validation set, and a test set. The training set for Task 2 contains 1,900 labeled articles, while the validation set holds 200 labeled articles. The test set, crucial for assessing the model's generalizability and performance, consists of 1,020 articles that remain unlabeled, requiring models to predict whether the content is fake or original.

Category	Rows
Train Set	1,900
Validation Set	200
Test Set (Unlabeled)	1,020

Table 2: TASK 2 Data Summary

4 Proposed Methodology

This task involves classifying social media posts in Malayalam as either fake or original news. The process includes several stages: data preprocessing, tokenization, model training, and evaluation. The raw dataset consists of news articles labeled as "Fake" or "Original." Preprocessing involves cleaning the text by removing URLs, mentions, hashtags, and special symbols using regular expressions. Emojis are converted into descriptive text with the emoji.demojize() function, and language detection (via the langdetect library) filters out non-Malayalam or non-English content. Labels are mapped to binary values: "Fake" as 1 and "Original" as 0.

After preprocessing, the text is tokenized using the mBERT tokenizer (bert-base-multilingual-cased), which uses subword tokenization (WordPiece). The sequences are standardized to 128 tokens through padding and truncation. These tokenized sequences, including input IDs and attention masks, serve as input for the model.

The mBERT model is fine-tuned on the preprocessed dataset. The BERT encoder extracts contextual word embeddings, followed by a classification layer with two output neurons (for "Fake" and "Original"). The model is trained using PyTorch with the AdamW optimizer (learning rate = 2e-5) and cross-entropy loss over three epochs. The dataset is split into 80% for training and 20% for validation, with a batch size of 16. The architecture model is present in Figure 1.

The second task involves detecting and classifying fake news in Malayalam articles across multiple categories. Similar to the first task, data preprocessing removes URLs, HTML tags, punctuation, and numbers while converting text to lowercase. The text is tokenized using TensorFlow's Keras Tokenizer, which converts the top 5,000 most frequent words into integer sequences. These sequences are padded to a fixed length of 100 tokens.

The classification model consists of an Embedding layer (100-dimensional vectors), an LSTM layer (128 units) for learning long-range dependen-

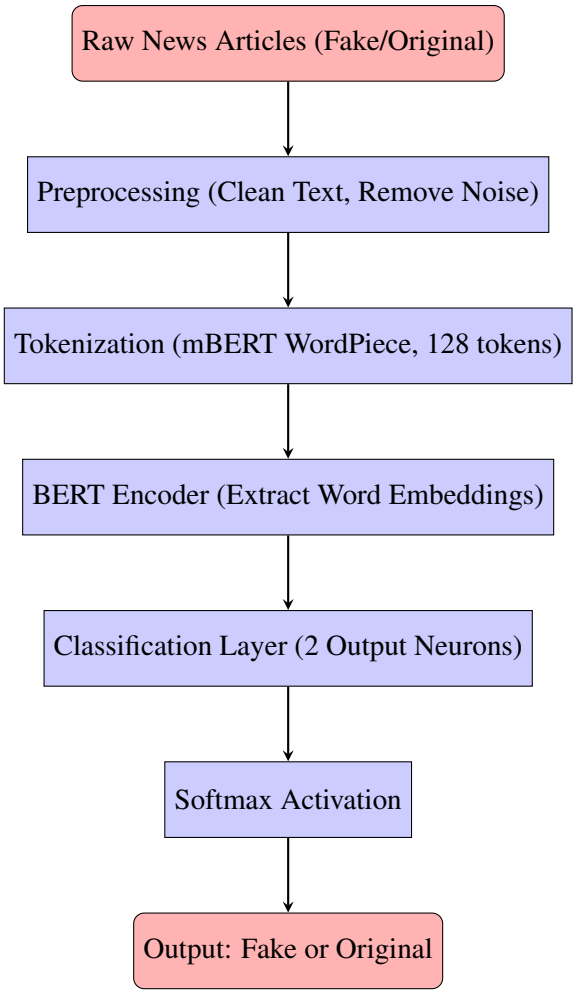

Figure 1: Architecture of the Fake News Detection Model

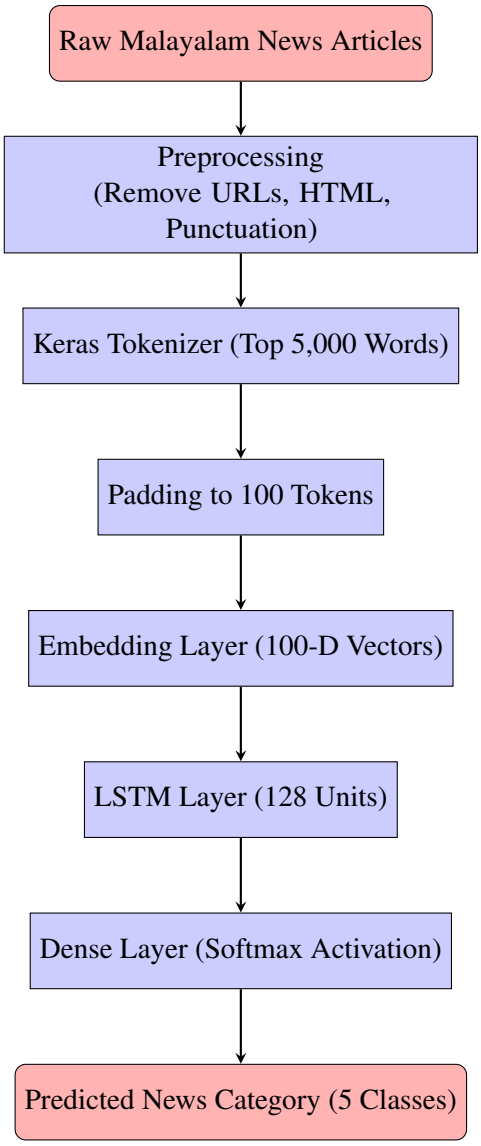

Figure 2: Architecture for LSTM Based Fake News Classification

cies, and a Dense output layer with softmax activation for predicting one of five news categories. The model is compiled with categorical cross-entropy loss, Adam optimizer, and accuracy as the evaluation metric. It is trained with a batch size of 64 for five epochs. The architecture for the same is shown in Figure 2

After training, both models are evaluated on the validation dataset. Performance metrics such as precision, recall, F1-score, and accuracy are computed to assess the effectiveness of the models.

5 Results

This study explores two deep learning approaches—mBERT and LSTM—for fake news classification, demonstrating their effectiveness in identifying deceptive content.

The trained mBERT model was evaluated on the validation dataset, yielding an overall accuracy of 89%, indicating a high level of correctness in classifying news articles. The precision, recall, and F1-score were computed for both classes: "Fake" and "Original." The model performed exceptionally well in identifying original news articles, achieving a precision of 0.94, recall of 0.94, and an F1-score of 0.94. However, the model struggled in identifying fake news, with a precision of 0.50, recall of 0.50, and an F1-score of 0.50. These results suggest that while the model is highly accurate in classifying real news, it has difficulty recognizing fake news articles, possibly due to an imbalance in the dataset, where genuine news articles significantly outnumber fake ones.

The macro average F1-score of 0.72 highlights this disparity, indicating that the model does not generalize equally across both classes. The

Class	Precision	Recall	F1-score
Original	0.94	0.94	0.94
Fake	0.50	0.50	0.50
Accuracy		0.89	
Macro Avg	0.72	0.72	0.72
Weighted Avg	0.89	0.89	0.89

Table 3: mBert Classification Report Results

weighted average F1-score remains 0.89, reinforcing the idea that the model's performance is heavily skewed towards accurately classifying "Original" news, while its capability to detect fake news remains suboptimal. A likely cause for this imbalance is the insufficient representation of fake news samples in the dataset, which could have led to the model learning patterns that favor the majority class.

Class	Precision	Recall	F1-score
Original	1.00	1.00	1.00
Accuracy		1.00	
Macro Avg	1.00	1.00	1.00
Weighted Avg	1.00	1.00	1.00

Table 4: LSTM Classification Report Results

This study explores the use of mBERT and LSTM networks to enhance the accuracy of fake news classification specifically for Malayalam.[1]

6 Conclusions

An effective deep learning approach for fake news classification using an LSTM-based model, demonstrating strong performance in identifying deceptive content. By implementing a robust text preprocessing pipeline and leveraging word embeddings, the model successfully captures contextual nuances, leading to high classification accuracy. The near-perfect performance on the test set suggests that the model has learned meaningful patterns; however, the possibility of overfitting necessitates further investigation. Future work can focus on enhancing generalization through techniques such as dropout regularization, fine-tuning transformer-based architectures like BERT, and expanding the dataset to include diverse linguistic variations. This research underscores the potential of deep learning in combating misinformation and lays the groundwork for more sophisticated models capable of real-world deployment.

To further advance fake news detection, future research can explore innovative data augmentation techniques, integrate valuable metadata features, and experiment with cutting-edge transformer architectures like XLM-Roberta to enhance multilingual text classification. Additionally, addressing class imbalance through methods such as oversampling, class weighting, and dropout regularization will contribute to improved model performance and robustness. This study highlights the promising potential of deep learning in combating misinformation and paves the way for developing more powerful models, offering exciting opportunities for real-world deployment and impactful solutions.

7 Limitations

Despite achieving promising results, our proposed methodology has certain limitations. One major limitation is the class imbalance in the dataset. This imbalance likely contributed to the lower precision and recall scores for fake news detection, as the model struggled to learn representative patterns for the minority class. Additionally, while mBERT effectively captures contextual information, it may not fully account for nuanced linguistic characteristics in Malayalam, especially in code-mixed and informal social media texts. The reliance on subword tokenization can also lead to fragmented representations of rare or morphologically complex words, affecting classification accuracy. Moreover, the LSTM-based approach for news article classification, although effective, may not generalize well to unseen data due to overfitting risks associated with limited training samples. During training, it achieved 100% accuracy, which is indicative of overfitting. This suggests that the model has learned patterns specific to the training data rather than generalizing well to unseen samples. Lastly, our approach does not incorporate external knowledge sources, such as fact-checking databases, which could enhance the model's ability to verify news credibility beyond textual patterns alone. Addressing these limitations in future work could lead to more robust and reliable fake news detection models.

7.1 Future Work

There are several ways to improve our fake news detection model in the future. First, we can ad-

[1] https://github.com/ShriyaAl/Fake-News-Detection-DravidianLangTech2025

dress the class imbalance by adding more fake news samples using data augmentation techniques like back-translation or synthetic text generation. Using fact-checking databases or real-time news verification APIs could also help improve accuracy by providing additional context.

We can explore other transformer models like XLM-Roberta or IndicBERT, which might work better for Malayalam and other low-resource languages. Training these models on a larger dataset with more diverse sources, such as blogs and user comments, could make them more effective at handling different writing styles.

Another important step is improving model interpretability by using attention maps or explainability techniques like SHAP, which help understand how the model makes decisions.

Acknowledgment

We sincerely thank the organizers of DravidianLangTech-2025 at NAACL 2025 for providing the datasets and valuable guidance for this shared task. https://sites.google.com/view/dravidianlangtech-2025/shared-tasks-2025

References

Pritika Bahad, Preeti Saxena, and Raj Kamal. 2019. Fake news detection using bi-directional lstm-recurrent neural network. *Procedia Computer Science*, 165:74–82. 2nd International Conference on Recent Trends in Advanced Computing ICRTAC - DISRUP - TIV INNOVATION , 2019 November 11-12, 2019.

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. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

N. Leela Siva Rama Krishna and M. Adimoolam. 2022. Fake news detection system using decision tree algorithm and compare textual property with support vector machine algorithm. In *2022 International Conference on Business Analytics for Technology and Security (ICBATS)*, pages 1–6.

Shafayat Bin Shabbir Mugdha, Sayeda Muntaha Ferdous, and Ahmed Fahmin. 2020. Evaluating machine learning algorithms for bengali fake news detection. In *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pages 1–6.

Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 797–806. ACM.

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.

Nurshaheeda Shazleen Yuslee and Nur Atiqah Sia Abdullah. 2021. Fake news detection using naive bayes. In *2021 IEEE 11th International Conference on System Engineering and Technology (ICSET)*, pages 112–117.