byteSizedLLM@DravidianLangTech 2025: Fake News Detection in Dravidian Languages Using Transliteration-Aware XLM-RoBERTa and Attention-BiLSTM

Rohith Gowtham Kodali
ASRlytics
Hyderabad, India
mdp0999@gmail.com

Durga Prasad Manukonda
ASRlytics
Hyderabad, India
rohitkodali@gmail.com

Abstract

This research introduces an innovative Attention BiLSTM-XLM-RoBERTa model for tackling the challenge of fake news detection in Malayalam datasets. By fine-tuning XLM-RoBERTa with Masked Language Modeling (MLM) on transliteration-aware data, the model effectively bridges linguistic and script diversity, seamlessly integrating native, Romanized, and mixed-script text. Although most of the training data is monolingual, the proposed approach demonstrates robust performance in handling diverse script variations. Achieving a macro F1-score of 0.5775 and securing top rankings in the shared task, this work highlights the potential of multilingual models in addressing resource-scarce language challenges and sets a foundation for future advancements in fake news detection.

1 Introduction

The rapid growth of social media platforms has revolutionized communication, enabling seamless information exchange and real-time updates. However, this connectivity has also fueled the spread of misinformation, or fake news. Detecting fake news has become a pressing challenge, particularly in resource-scarce languages like Malayalam.

The Fake News Detection in Dravidian Languages - DravidianLangTech@NAACL 2025[1] (Subramanian et al., 2025, 2023, 2024b) shared task provides a platform for researchers to tackle the critical challenge of detecting fake news in Malayalam-language news articles. Task 2, the FakeDetect-Malayalam shared task, focuses on classifying misinformation into five nuanced categories. In an age of information overload, accurate detection is crucial for fostering trustworthy communication and curbing the spread of misinformation. The task seeks to inspire the development of effective models designed to address the unique linguistic and contextual complexities of Malayalam.

Our study presents a robust architecture combining fine-tuned XLM-RoBERTa embeddings with a custom Attention-BiLSTM classifier to enhance contextual understanding and capture complex sequential dependencies in multilingual text. The embeddings, trained using Masked Language Modeling (MLM), were derived from the AI4Bharath dataset, incorporating diverse transliteration patterns to handle linguistic and orthographic variability. This approach enables effective processing of native scripts, Romanized text, and mixed-script data. Despite the monolingual dominance in training data, the model outperforms baselines, demonstrating strong cross-lingual adaptability. The Attention-BiLSTM classifier, leveraging general attention mechanisms, ensures precise classification in complex linguistic scenarios.

This study analyzes data preprocessing, MLM training, and classifier design, introducing innovations for improved accuracy and scalability. It establishes a robust framework for fake news detection in Dravidian languages, offering insights into model performance and deployment challenges.

2 Related Work

The growing challenge of disinformation has driven extensive research into fake news detection. Raja et al. (2023) explored detecting fake news in Dravidian languages using transfer learning with adaptive fine-tuning, while Keya et al. (2022) employed a pretrained BERT model with data augmentation, benchmarking its performance against other models. Similarly, Goldani et al. (2021) investigated capsule networks for extracting n-gram-based features.

Research efforts have also addressed fake news detection in low-resource languages. Gereme,

[1] https://codalab.lisn.upsaclay.fr/competitions/20698

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 182–187
May 3, 2025 ©2025 Association for Computational Linguistics

Fantahun and Zhu, William and Ayall, Tewodros and Alemu, Dagmawi (2021) and Saghayan et al. (2021) focused on Amharic and Persian, respectively, while Faustini and Covões (2020) emphasized the importance of addressing fake news in resource-poor languages, including Dravidian languages. Furthermore, Vijjali et al. (2020) proposed a two-stage pipeline leveraging BERT and ALBERT for detecting COVID-19-related misinformation.

The shared tasks on Fake News Detection in Malayalam, organized by DravidianLangTech@EACL 2023 (S et al., 2023; Subramanian et al., 2023) and 2024 (Subramanian et al., 2024a,b), focused on classifying fake news, low-resource settings. The top-performing teams in the 2024 challenge utilized pre-trained Malayalam BERT (Rahman et al., 2024; Tabassum et al., 2024), and XLM-RoBERTa Base (Osama et al., 2024) models, while in 2023, they relied on XLM-RoBERTa (Luo and Wang, 2023), and MuRIL (Bala and Krishnamurthy, 2023) models. These tasks underscored challenges with transliterated and mixed-script data, highlighting the need for robust training and fine-tuned LLMs like XLM-RoBERTa, MuRIL and BERT, which effectively handle linguistic nuances for accurate fake news detection.

3 Dataset

The **Fake News Detection from Malayalam News (FakeDetect-Malayalam)** shared task focuses on detecting and classifying fake news in Malayalam-language news articles. Accurate detection is critical for mitigating misinformation and ensuring reliable communication. Task 2 involves classifying news articles into five categories: *False*, *Half True*, *Mostly False*, *Partly False*, and *Mostly True* (Devika et al., 2024).

The dataset comprises social media comments and news articles, annotated for these categories. It is split into training and testing sets to ensure balanced distribution, as shown in Table 1.

This dataset forms a strong foundation for training models to handle the linguistic and contextual nuances of Malayalam, advancing fake news detection in low-resource settings.

Label	**Train**	**Test**	**Total**
FALSE	1386	100	1486
MOSTLY FALSE	295	56	351
HALF TRUE	162	37	199
PARTLY FALSE	57	7	64
Total	1900	200	2100

Table 1: Dataset distribution for Task 2: Fake news detection in Malayalam.

4 Methodology

This section presents our proposed architecture, combining fine-tuned XLM-RoBERTa embeddings with an Attention BiLSTM classifier. The following subsections detail our approach.

4.1 Fine-Tuning XLM-RoBERTa with MLM

XLM-RoBERTa, a multilingual transformer model trained on 94 languages (Conneau et al., 2019), was fine-tuned using Masked Language Modeling (MLM) to enhance its contextual embeddings for both multilingual and monolingual Malayalam text. MLM involves masking portions of input text and training the model to predict them, enabling it to learn representations suited to the linguistic and script challenges of Malayalam.

The fine-tuning dataset included monolingual Malayalam text, fully Romanized transliterations, and mixed-script data with 20–70% transliterated words per sentence. This approach enabled the model to effectively handle native scripts, Romanized text, and mixed-script variations commonly found in Malayalam social media. The fine-tuned XLM-RoBERTa model[2] serves as a robust embedding backbone, addressing both multilingual and monolingual linguistic variability in Malayalam text.

4.2 Attention BiLSTM-XLM-RoBERTa Model

This study proposes a hybrid Attention BiLSTM-XLM-RoBERTa model (Liu and Guo, 2019; Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005; Kodali et al., 2025; Manukonda and Kodali, 2025, 2024a; Kodali and Manukonda, 2024; Manukonda and Kodali, 2024b) for multi-label classification. As illustrated in Figure 1, the model integrates fine-tuned XLM-RoBERTa embeddings with a BiLSTM and attention mechanism to capture rich language-specific features.

The input sequence is passed through XLM-RoBERTa to generate contextual embeddings $\mathbf{X} \in$

[2] https://huggingface.co/bytesizedllm/MalayalamXLM_Roberta

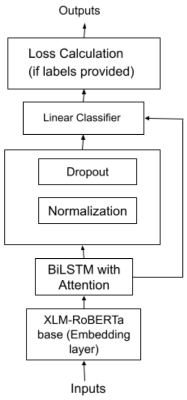

Figure 1: Architecture of the BiLSTM-XLM-RoBERTa Classifier Model. Residual components like layer normalization and dropout regularization enhance generalization.

$R^{T\times 768}$:

$$\mathbf{X} = \mathbf{XLMRoBERTa}(input\_ids, attention\_mask) \quad (1)$$

These embeddings are processed by a BiLSTM, which produces forward and backward hidden states $\mathbf{H}_{fwd}$ and $\mathbf{H}_{bwd}$. The combined hidden state at each time step t is:

$$\mathbf{H}_t = [\mathbf{H}_{fwd,t}; \mathbf{H}_{bwd,t}] \quad (2)$$

An attention mechanism assigns importance to each hidden state, generating attention weights α_t:

$$\mathbf{a}_t = \tanh(\mathbf{W}_{att} \cdot \mathbf{H}_t), \quad \alpha_t = \frac{\exp(\mathbf{a}_t)}{\sum_{t=1}^{T} \exp(\mathbf{a}_t)} \quad (3)$$

The attention-weighted representation is computed as:

$$\mathbf{H}_{attended} = \sum_{t=1}^{T} \alpha_t \cdot \mathbf{H}_t \quad (4)$$

Residual components such as layer normalization and dropout are applied to the attention-weighted representation to stabilize training and reduce overfitting:

$$\mathbf{H}_{dropout} = Dropout(LayerNorm(\mathbf{H}_{attended})) \quad (5)$$

Finally, a classification layer outputs logits:

$$\mathbf{logits} = \mathbf{W}_{cls} \cdot \mathbf{H}_{dropout} + \mathbf{b}_{cls} \quad (6)$$

The model is trained using cross-entropy loss:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad (7)$$

This architecture effectively combines fine tuned XLM-RoBERTa base embeddings, BiLSTM processing, and attention to enhance multi-label classification performance.

5 Experiment Setup

The experiment employed transliteration-aware fine-tuning for Malayalam fake news detection by combining XLM-RoBERTa fine-tuning with MLM and integrating embeddings into an Attention-BiLSTM classifier.

5.1 Fine-Tuning the XLM-RoBERTa Model

XLM-RoBERTa was fine-tuned using MLM on a transliteration-aware dataset derived from 340MB of Malayalam monolingual text sourced from AI4Bharath (Kunchukuttan et al., 2020). Using IndicTrans (Bhat et al., 2015), the dataset was transformed into three variants: original Malayalam script, fully transliterated Roman script, and partially transliterated text with 20–70% transliterated words per sentence. This ensured exposure to transliteration patterns and orthographic variations common in social media.

Fine-tuning used a 9:1 train-validation split, a 15% masking probability, a batch size of 16, and a learning rate of 5×10^{-5}. Training ran for up to 10 epochs, with early stopping based on validation perplexity to optimize embeddings for mixed-script Malayalam text.

5.2 Integration into Attention BiLSTM

The fine-tuned embeddings, ‘MalayalamXLM_Roberta‘, were input into an Attention BiLSTM classifier with an input size of 768, a hidden size of 512, and 3 LSTM layers. The attention mechanism captured critical features and dependencies in multilingual sequences.

Dropout (0.5) and layer normalization were applied to stabilize training and reduce overfitting. The AdamW optimizer with a learning rate of 1×10^{-5} was used, with early stopping based on validation loss and macro F1-score ensuring robust performance.

This transliteration-aware MLM fine-tuning and Attention BiLSTM setup effectively handled transliterated and mixed-script Malayalam text.

Label	Precision	Recall	F1-Score	Support
FALSE	0.67	0.83	0.74	100
HALF TRUE	0.48	0.30	0.37	37
PARTLY FALSE	1.00	0.57	0.73	7
MOSTLY FALSE	0.50	0.45	0.47	56
Accuracy	-	-	0.61	200
Macro Avg	0.66	0.54	0.58	200
Weighted Avg	0.60	0.61	0.60	200

Table 2: Classification Report on the Test Set for Fake News Detection

Team Name	mF1	Rank
KCRL	0.6283	1
byteSizedLLM	**0.5775**	**2**
NLP_goats	0.5417	4

Table 3: Macro F1 (mF1) scores and ranks of top3 performing teams.

6 Results and Discussion

The proposed Attention BiLSTM-XLM-RoBERTa model demonstrated competitive performance in fake news detection on the Malayalam-English code-mixed dataset[3]. As shown in Table 2, the model achieved an overall accuracy of 61% with a macro F1-score of 0.58. The 'FALSE' label exhibited the highest F1-score of 0.74, while the 'HALF TRUE' label scored the lowest at 0.37, reflecting challenges posed by imbalanced data.

The fine-tuned MalayalamXLM_Roberta model, optimized with Masked Language Modeling (MLM), achieved a perplexity of 4.15, generating effective contextual embeddings. When used independently, these embeddings achieved a macro F1-score of 0.5394. Integrating them into the Attention BiLSTM classifier improved performance to a macro F1-score of 0.5775 with an optimal configuration of a learning rate of 1×10^{-5}, an LSTM hidden size of 512, and 3 LSTM layers. Other configurations, such as a learning rate of 2×10^{-5} with 256 hidden units and 2 LSTM layers, resulted in a slightly lower F1-score of 0.5718. Comparatively, an advanced encoder-decoder transformer model achieved a macro F1-score of 0.5532, reaffirming the efficiency of the Attention BiLSTM approach for small datasets.

As shown in Table 3, our team, **ByteSizedLLM**, secured second and third ranks in the shared task with macro F1-scores of 0.5775 and 0.5718, respectively. Despite most of the training data being monolingual, the multilingual XLM-RoBERTa model exhibited remarkable robustness in handling code-mixed scenarios, highlighting its adaptability across diverse linguistic contexts.

[3] https://github.com/mdp0999/Fake-News-Detection/blob/main/task2.ipynb

7 Limitations and Future Work

The model's performance was limited by the dataset size, which was restricted to 340MB of code-mixed text due to computational constraints. Additionally, inaccuracies in the transliteration process may have impacted the quality of embeddings. The imbalanced label distribution also posed challenges, particularly for minority classes like 'HALF TRUE' and 'PARTLY FALSE'.

Future work aims to overcome limitations by using larger datasets, improving transliteration, and exploring advanced architectures for better fake news detection in multilingual and code-mixed contexts.

8 Conclusion

This study proposed an Attention BiLSTM-XLM-RoBERTa model for fake news detection in Malayalam datasets. By fine-tuning XLM-RoBERTa with MLM on transliteration-aware data and integrating the embeddings into an attention-enhanced BiLSTM architecture, the approach effectively addressed linguistic and script challenges in Malayalam text. The model achieved a macro F1-score of 0.5775, securing top rankings in the shared task and demonstrating its robustness in resource-constrained settings.

Despite the predominantly monolingual nature of the training data and transliteration limitations, the model performed strongly, showcasing the ability of multilingual XLM-RoBERTa embeddings to handle diverse script variations. These results underscore the potential of multilingual models for low-resource language tasks.

References

Abhinaba Bala and Parameswari Krishnamurthy. 2023. AbhiPaw@DravidianLangTech: Multimodal abusive language detection and sentiment analysis. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 140–146, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2015. Iiit-h system submission for fire2014 shared task on transliterated search. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '14, pages 48–53, New York, NY, USA. ACM.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

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.

Pedro Henrique Arruda Faustini and Thiago Ferreira Covões. 2020. Fake news detection in multiple platforms and languages. *Expert Systems with Applications*, 158:113503.

Gereme, Fantahun and Zhu, William and Ayall, Tewodros and Alemu, Dagmawi. 2021. Combating fake news in "low-resource" languages: Amharic fake news detection accompanied by resource crafting. *Information*, 12(1).

Mohammad Hadi Goldani, Saeedeh Momtazi, and Reza Safabakhsh. 2021. Detecting fake news with capsule neural networks. *Applied Soft Computing*, 101:106991.

A. Graves and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052 vol. 4.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Ashfia Jannat Keya, Md. Anwar Hussen Wadud, M. F. Mridha, Mohammed Alatiyyah, and Md. Abdul Hamid. 2022. AugFake-BERT: Handling Imbalance through Augmentation of Fake News Using BERT to Enhance the Performance of Fake News Classification. *Applied Sciences*, 12(17).

Rohith Kodali and Durga Manukonda. 2024. byteSizedLLM@DravidianLangTech 2024: Fake news detection in Dravidian languages - unleashing the power of custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 79–84, St. Julian's, Malta. Association for Computational Linguistics.

Rohith Gowtham Kodali, Durga Prasad Manukonda, and Daniel Iglesias. 2025. byteSizedLLM@NLU of Devanagari script languages 2025: Hate speech detection and target identification using customized attention BiLSTM and XLM-RoBERTa base embeddings. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 242–247, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

Gang Liu and Jiabao Guo. 2019. Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337:325–338.

Zhipeng Luo and Jiahui Wang. 2023. DeepBlueAI@DravidianLangTech-RANLP 2023. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 171–175, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Durga Manukonda and Rohith Kodali. 2024a. byteLLM@LT-EDI-2024: Homophobia/transphobia detection in social media comments - custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 157–163, St. Julian's, Malta. Association for Computational Linguistics.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2024b. Enhancing multilingual natural language processing with custom subword tokenization: Subword2vec and bilstm integration for lightweight and streamlined approaches. In *2024 6th International Conference on Natural Language Processing (ICNLP)*, pages 366–371.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2025. byteSizedLLM@NLU of Devanagari script languages 2025: Language identification using customized attention BiLSTM and XLM-RoBERTa base embeddings. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 248–252, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Md Osama, Kawsar Ahmed, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024.

CUET_NLP_GoodFellows@DravidianLangTech EACL2024: A transformer-based approach for detecting fake news in Dravidian languages. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 187–192, St. Julian's, Malta. 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.

Eduri Raja, Badal Soni, and Samir Kumar Borgohain. 2023. Fake news detection in Dravidian languages using transfer learning with adaptive finetuning. *Engineering Applications of Artificial Intelligence*, 126:106877.

Malliga S, Bharathi Raja Chakravarthi, Kogilavani S V, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, and Muskaan Singh. 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*, pages 59–63, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Masood Hamed Saghayan, Seyedeh Fatemeh Ebrahimi, and Mohammad Bahrani. 2021. Exploring the Impact of Machine Translation on Fake News Detection: A Case Study on Persian Tweets about COVID-19. In *2021 29th Iranian Conference on Electrical Engineering (ICEE)*, pages 540–544.

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, Premjith B, Vanaja K, Mithunja S, Devika K, Hariprasath S.b, Haripriya B, and Vigneshwar E. 2024a. 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, St. Julian's, Malta. 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. 2024b. 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.

Nafisa Tabassum, Sumaiya Aodhora, Rowshon Akter, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. Punny_Punctuators@DravidianLangTech-EACL2024: Transformer-based approach for detection and classification of fake news in Malayalam social media text. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 180–186, St. Julian's, Malta. Association for Computational Linguistics.

Rutvik Vijjali, Prathyush Potluri, Siddharth Kumar, and Sundeep Teki. 2020. Two Stage Transformer Model for COVID-19 Fake News Detection and Fact Checking. *Preprint*, arXiv:2011.13253.