byteSizedLLM@DravidianLangTech 2025: Fake News Detection in Dravidian Languages Using Transliteration-Aware XLM-RoBERTa and Transformer Encoder-Decoder

Durga Prasad Manukonda
ASRlytics
Hyderabad, India
mdp0999@gmail.com

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

Abstract

This study addresses the challenge of fake news detection in code-mixed and transliterated text, focusing on a multilingual setting with significant linguistic variability. A novel approach is proposed, leveraging a fine-tuned multilingual transformer model trained using Masked Language Modeling on a dataset that includes original, fully transliterated, and partially transliterated text. The fine-tuned embeddings are integrated into a custom transformer classifier designed to capture complex dependencies in multilingual sequences. The system achieves state-of-the-art performance, demonstrating the effectiveness of combining transliteration-aware fine-tuning with robust transformer architectures to handle code-mixed and resource-scarce text, providing a scalable solution for multilingual natural language processing tasks.

1 Introduction

The rise of social media platforms like Facebook, X (formerly Twitter), and Instagram has revolutionized global connectivity, enabling instant information sharing. However, it has also fueled the spread of fake news—intentionally misleading content—causing societal issues such as eroded media trust, polarized opinions, and real-world consequences. Addressing fake news detection is now a critical research area (Subramanian et al., 2023, 2024b).

This study focuses on Task 1 of the shared challenge, Fake News Detection in Dravidian Languages - DravidianLangTech@NAACL 2025 (Subramanian et al., 2025), which classifies social media posts as original or fake. Unlike traditional news, social media content is user-generated, informal, and diverse in style, making fake news detection particularly complex. The goal is to develop a robust classification system using advanced computational techniques and machine learning models.

To tackle multilingual challenges, we introduce the TransformerXLMRoberta Classifier, a hybrid model that utilizes fine-tuned XLM-RoBERTa with Masked Language Modeling (MLM) on original, fully, and partially transliterated datasets. This enables handling of native scripts, Romanized text, and mixed-script data. Additionally, fine-tuned XLM-RoBERTa embeddings are enhanced through a hybrid architecture with a custom transformer design, projected to match transformer dimensions, and refined via Encoder-Decoder layers to capture complex contextual relationships. Regularization techniques such as dropout and gradient clipping ensure stable training.

This approach achieves state-of-the-art performance in multilingual text classification, highlighting the role of transliteration strategies and hybrid architectures in addressing the challenges of multilingual and transliterated data. By advancing NLP for resource-scarce languages, this work contributes to more inclusive and effective multilingual applications.

2 Related Work

The rising prevalence of disinformation has driven significant 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) utilized a pretrained BERT model with data augmentation, comparing results across multiple models. Similarly, Goldani et al. (2021) investigated capsule networks for n-gram-based feature extraction.

Beyond English, Gereme, Fantahun and Zhu, William and Ayall, Tewodros and Alemu, Dagmawi (2021) and Saghayan et al. (2021) examined fake news detection in Amharic and Persian. Chu et al. (2021) demonstrated the cross-lingual effectiveness of BERT, while Faustini and Covões (2020) emphasized resource-poor languages, including Dra-

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 176–181
May 3, 2025 ©2025 Association for Computational Linguistics

vidian languages. Vijjali et al. (2020) proposed a two-stage pipeline using BERT and ALBERT for verifying COVID-19 fake news.

The Fake News Detection in Malayalam - DravidianLangTech@EACL 2023 (S et al., 2023) and 2024 (Subramanian et al., 2024a) shared tasks focused on classifying fake news in low-resource settings, addressing transliteration and mixed-script challenges. 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 highlighted the effectiveness of multilingual models like XLM-RoBERTa, MuRIL and BERT in improving fake news detection across diverse linguistic contexts.

3 Dataset

The dataset for **Task 1** of the shared task *"Fake News Detection in Dravidian Languages - DravidianLangTech@NAACL 2025"* (Devika et al., 2024) consists of social media posts from platforms such as Twitter, Facebook, and YouTube. These posts are categorized as either *fake* or *original*. The dataset is divided into three splits: training, development, and testing, ensuring a balanced distribution for robust evaluation.

The data distribution across the splits is summarized in Table 1.

Dataset Split	Fake	Original	Total
Train	1,599	1,658	3,257
Development(Dev)	406	409	815
Test	507	512	1,019

Table 1: Data distribution for Fake News Detection in Dravidian Languages Task 1

The dataset reflects real-world challenges in fake news detection by including posts with informal language, transliterated text, and mixed-script content. Participants are tasked with designing systems to classify each post or comment as either *fake* or *original*, providing a benchmark for robust and multilingual fake news detection systems.

4 Methodology

This section introduces our proposed architecture, which integrates fine-tuned XLM-RoBERTa embeddings with a robust Transformer-based classifier.

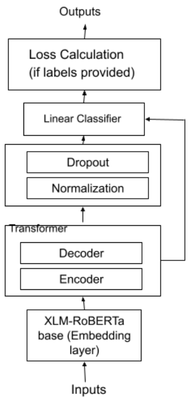

Figure 1: Architecture of the Custom Transformer XLM-Roberta Classifier Model.

The fine-tuned embeddings, trained using Masked Language Modeling (MLM), enhance contextual understanding, while the classifier captures complex sequential dependencies in multilingual and transliterated text. The following subsections detail the data preprocessing, MLM training, and classifier design.

4.1 XLM-RoBERTa Base Fine-Tuned with MLM

XLM-RoBERTa, a multilingual transformer model trained on a large-scale corpus of 94 languages (Conneau et al., 2019), was fine-tuned using Masked Language Modeling (MLM) for this study. MLM involves masking a subset of input tokens and training the model to predict them, allowing it to learn enriched contextual embeddings tailored to the bilingual challenges of Malayalam-English datasets.

The MLM training dataset included monolingual text from Malayalam social media sources, fully transliterated versions of this text in Roman script, and partially transliterated data where 20–70% of words in each sentence were transliterated. This strategy enabled the model to handle native scripts, Romanized text, and mixed-script text commonly found in social media communication. The fine-tuned XLM-RoBERTa model [1] serves as the embedding backbone for downstream classification tasks, effectively addressing linguistic and orthographic variability in multilingual datasets.

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

	Precision	**Recall**	**F1-Score**	**Support**
original	0.89	0.90	0.90	512
Fake	0.90	0.89	0.90	507
Macro Avg	0.90	0.90	0.90	1019
Weighted Avg	0.90	0.90	0.90	1019
Accuracy	-	-	0.90	1019

Table 2: Classification Report on the Test Set

4.2 Custom Transformer XLMRoberta Classifier

The proposed custom transformer architecture called TransformerXLMRobertaClassifier, integrates XLM-RoBERTa embeddings with a transformer-based encoder-decoder design to effectively handle multilingual and code-mixed text, drawing on the foundational Transformer architecture (Vaswani et al., 2023) and inspired by our prior research on architecture design (Manukonda and Kodali, 2025; Kodali et al., 2025; Kodali and Manukonda, 2024; Manukonda and Kodali, 2024a,b). The model begins by processing input token IDs and attention masks through the fine-tuned XLM-RoBERTa model to generate contextual embeddings. These embeddings are then projected into the transformer's input dimension and passed through encoder and decoder layers, utilizing attention mechanisms and masking to capture complex dependencies across sequences.

The decoder outputs are aggregated into a fixed-dimensional representation and refined with residual layers, normalization, and dropout for enhanced generalization. The final output is passed through a classification layer to produce logits, with cross-entropy loss computed during supervised training.

By combining XLM-RoBERTa embeddings with transformer-based attention mechanisms, the TransformerXLMRobertaClassifier effectively addresses the challenges of multilingual and transliterated text, ensuring robust and efficient performance. As illustrated in Figure 1, the architecture leverages regularization techniques such as dropout and masking to maintain stability and prevent overfitting.

5 Experiment Setup

The experiment setup involved transliteration-aware fine-tuning for fake news detection in Malayalam-English code-mixed datasets, comprising XLM-RoBERTa fine-tuning with Masked Language Modeling (MLM) and embedding integration into a custom transformer-based classifier.

5.1 Fine-Tuning the XLM-RoBERTa Model

XLM-RoBERTa was fine-tuned using masked language modeling (MLM) with a transliteration-aware strategy on a 340MB Malayalam-English code-mixed dataset from AI4Bharath (Kunchukuttan et al., 2020), prepared using IndicTrans (Bhat et al., 2015). The dataset included three text variants: Malayalam script, fully transliterated Roman script, and partially transliterated text, exposing the model to diverse transliteration patterns in social media communication.

The data was split 9:1 for training and validation. Fine-tuning used a 15% masking probability, batch size 16, and a 5×10^{-5} learning rate for up to 10 epochs on GPUs, with early stopping based on validation perplexity to prevent overfitting. The fine-tuned embeddings optimized handling of transliterated and mixed-script text.

5.2 Integration into Custom Transformer Classifier

The fine-tuned 'MalayalamXLM_Roberta' model demonstrated effectiveness in capturing transliteration patterns. These embeddings were integrated into 'TransformerXLMRobertaClassifier', a custom transformer classifier with three encoder-decoder layers, hidden dimension 768, 8 attention heads, and a 2048 feedforward dimension. Attention mechanisms captured multilingual dependencies effectively.

Dropout (0.3) and normalization were applied in residual layers to enhance generalization. AdamW optimizer with a 1×10^{-5} learning rate was used, with early stopping based on validation loss and macro F1-score.

This two-stage approach—transliteration-aware MLM fine-tuning followed by transformer-based classification—effectively addressed Malayalam-English code-mixed and transliterated text challenges.

Model	F1 Macro
XLM-RoBERTa Base	0.8675
MalayalamXLM_Roberta (Fine-Tuned MLM)	0.8900
Attention-BiLSTM MalayalamXLM_Roberta	0.8969
TransformerXLMRobertaClassifier (Proposed)	**0.8979**

Table 3: Macro F1 scores for various models on Malayalam-English code-mixed fake news detection.

5.3 Evaluation

The models were evaluated using macro F1-score, accuracy, and perplexity. Macro F1-score addressed class imbalance, accuracy measured overall correctness, and perplexity assessed the model's ability to predict masked tokens, with lower values indicating better adaptation.

6 Results and Discussion

The fine-tuned MalayalamXLM_Roberta model achieved a perplexity score of 4.1, showcasing its effectiveness in capturing transliteration patterns.

Table 3 summarizes the performance of various models on the Malayalam-English fake news detection task. The base XLM-RoBERTa achieved a macro F1-score of 0.8675. Fine-tuning with MLM improved this to 0.8900 with MalayalamXLM_Roberta. Adding attention mechanisms in the Attention-BiLSTM MalayalamXLM_Roberta model raised the score to 0.8969. The proposed TransformerXLMRobertaClassifier[2] achieved the highest macro F1-score of **0.8979**, highlighting the effectiveness of transliteration-aware fine-tuning and the custom architecture.

The success of this approach was further demonstrated in the shared task results, where our team, **bytesizedllm**, achieved the highest macro F1-score of **0.8979 (0.898)**. A detailed analysis of the test set results is provided in Table 2, and our team secured the top position among all participating teams. Table 4 highlights the rankings and comparative scores of the top-performing teams.

Team Name	mF1	Rank
bytesizedllm	**0.898**	**1**
CUET_NLP_MP_MD	0.893	2
One_by_zero	0.892	3

Table 4: Macro F1 (mF1) scores and ranks of top3 teams.

The results underscore the importance of transliteration-aware fine-tuning in addressing the complexities of code-mixed and multilingual text. By incorporating fully and partially transliterated datasets, the models demonstrated robust generalization across native scripts, Romanized text, and mixed-script patterns. The 'TransformerXLMRobertaClassifier' further amplified these gains by capturing dependencies effectively through its custom architecture.

7 Limitations and Future Work

The model's performance was limited by the dataset size, which was restricted to a small of code-mixed text due to computational constraints. Additionally, inaccuracies in the transliteration process may have impacted the quality of embeddings.

Future work will address these limitations by training on larger datasets, refining transliteration, and exploring advanced architectures to enhance fake news detection in multilingual and code-mixed contexts.

8 Conclusion

This study proposes a transliteration-aware fine-tuning approach for fake news detection in Malayalam-English code-mixed text. By fine-tuning XLM-RoBERTa on fully and partially transliterated datasets and integrating the resulting embeddings into a custom transformer classifier, the method demonstrated state-of-the-art performance.

The custom transformer model, Transformer XLMRoberta Classifier, consistently outperformed baseline models, highlighting the effectiveness of combining transliteration-aware pretraining with advanced architectures. These findings contribute significantly to the advancement of multilingual NLP, providing a robust framework for tackling the complexities of code-mixed and resource-scarce languages like Malayalam.

[2] https://github.com/mdp0999/Fake-News-Detection/blob/main/task1_m.ipynb

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.

Samuel Kai Wah Chu, Runbin Xie, and Yanshu Wang. 2021. Cross-Language Fake News Detection. *Data and Information Management*, 5(1):100–109.

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.

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*.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

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.