Celestia@DravidianLangTech 2025: Malayalam-BERT and m-BERT based transformer models for Fake News Detection in Dravidian Languages

Syeda Alisha Noor[1], Sadia Anjum[2], Syed Ahmad Reza[1], and Md. Rashadur Rahman[1]

[1]Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
[2]International Islamic University Chittagong, Chattogram, Bangladesh
u1904018@student.cuet.ac.bd, sadiaanjum210@gmail.com, u1904016@student.cuet.ac.bd
rashadur@cuet.ac.bd

Abstract

Fake news detection in Malayalam is difficult due to limited data and language challenges. This study compares machine learning, deep learning, and transformer models for classification. The dataset is balanced and divided into training, development and test sets. Machine learning models (SVM, Random Forest, Naive Bayes) used TF-IDF features and deep learning models (LSTM, BiLSTM, CNN) worked with tokenized sequences. We fine-tuned transformer models like IndicBERT, MuRIL, mBERT, and Malayalam-Bert. Among them, the Malayalam-Bert model performed the best and achieved an F1 score of 86%. On the other hand mBERT performed best at spotting fake news. However, the models struggled with mixed-language text and complex writing. Despite these challenges, transformer models turned out to be the most effective for detecting fake news in Malayalam.

1 Introduction

Fake news is spreading fast on the internet, and it has become important to find better ways to detect it, especially in languages that do not have many digital resources. The shared task on fake news detection in Dravidian languages, held at DravidianLangTech@NAACL 2025, was organized to tackle this problem. This research focused on developing and testing different methods to identify fake news in Dravidian languages. Since these languages are complex, the task used special approaches instead of general models.

Fake news detection (FND) can be divided into two types: monolingual and multilingual. Monolingual FND is used to find fake news in one language. On the other hand, multilingual FND is needed when fake news is mixed with multiple languages, including code-mixed content. Detecting fake news in low-resource languages is difficult because there is a lack of enough labeled datasets, pre-trained models, or other digital tools. However, some methods like collecting and labeling data, using cross-lingual models, applying transfer learning, and creating models suited for specific languages can help to improve fake news detection.

In this study, we tested four pre-trained transformer models, Indic-BERT, m-BERT, Malayalam and MuRIL, to determine whether transfer learning from high-resource languages could improve fake news detection in Dravidian languages. Their findings demonstrated the feasibility of this approach, highlighting the role of advanced NLP techniques in mitigating misinformation in underrepresented languages. The implementation details have been provided in the following GitHub repository:- https://github.com/Alisha1904018/Share-task-2025.

2 Related Work

Fake news detection has become a growing area of research nowadays. Low-resource languages like Dravidian languages form a significant field of study. Many studies have focused on using machine learning and deep learning approaches to address this problem.

(Raja et al., 2024) developed a hybrid model that combines CNN and BiLSTM to detect fake news in Dravidian languages. They used MuRIL to obtain better language-specific details and reduce overfitting. Their model performed better than state-of-the-art approaches.

(Shanmugavadivel et al., 2024) also took part in DravidianLangTech 2024 and experimented with machine learning models such as Random Forest, Logistic Regression, and Decision Trees.

(Farsi et al., 2024) customized a MuRIL-BERT model and evaluated it with different machine learning and transformer-based techniques. Their strategy resulted in an F1-score of 0.86 for binary classification and 0.5191 for multi-class classification.

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 729–734

(Shohan et al., 2024)achieved the highest F1 scores of 75.82% by using RoBERTa for English tweets in classifying check worthy sentences.

(Rahman et al., 2024) achieved the highest macro F1-score (0.88) in Malayalam fake news detection using Malayalam-BERT, securing the top position in the shared task.

(Osama et al., 2024) explored both machine learning models, such as SVM, Random Forest, Logistic Regression, and Naïve Bayes, and deep learning models: CNN, BiLSTM, and BiLSTM with attention, along with transformers. The best-performing model in this study was m-BERT, which had an F1-score of 0.85 and ranked 4th in the shared task.

(Borgohain et al., 2023) created a dataset named Dravidian_Fake, containing 26,000 fake news articles in Telugu, Tamil, Kannada, and Malayalam languages. This study has attained the highest accuracy of 93.31% with mBERT and XLM-R using adaptive fine-tuning for multilingual fake news classification.

(Raja et al., 2023) tested four transformer models—mBERT, XLM-RoBERTa, IndicBERT, and MuRIL—on Telugu, Kannada, Tamil, and Malayalam fake news detection. Among these, MuRIL performed the best.

(Yigezu et al., 2024) introduced *Ethio-Fake*, a framework that integrates social-contextual and content-driven attributes for misinformation detection in low-resource languages. They evaluated various techniques, including traditional machine learning, neural networks, and transfer learning, concluding that ensemble learning achieved the highest F1-score of 0.99.

(Wang et al., 2024) conducted an extensive survey on monolingual and multilingual misinformation detection for under-resourced languages. Their work reviewed existing datasets, methodologies, and challenges in the field, emphasizing the need for improved data collection and inclusive AI strategies. They also highlighted the effectiveness of language-agnostic and multi-modal approaches in combating misinformation.

(Shimi et al., 2024) focused on *language identification* for Dravidian languages, a crucial step in fake news detection within multilingual settings. They compared machine learning and deep learning models for recognizing languages like Tamil and Malayalam. Their results indicated that deep learning-based language-independent models achieved the highest accuracy of 98%.

3 Task and Dataset Description

Fake News Detection in Dravidian Languages comprises balanced and normalized data (Subramanian et al., 2025) given by the organizers (Subramanian et al., 2024), basically aiming at building some systems (Devika et al., 2024) to label a original versus fake news of the Malayalam language (Subramanian et al., 2023) posts found in the media.The dataset consists of 3 portion: train, test & dev dataset

Class	Train	Development	Test
Fake	1599	406	507
Original	1658	409	512
Total	3257	815	519

Table 1: Dataset analysis

4 Methodology

The proposed method is experimented by using different machine learning, deep learning and transformer-based approach to classify fake news in a code-mixed Malayalam-English dataset. Our approach consists of data preprocessing, feature extraction, model development, and performance evaluation.

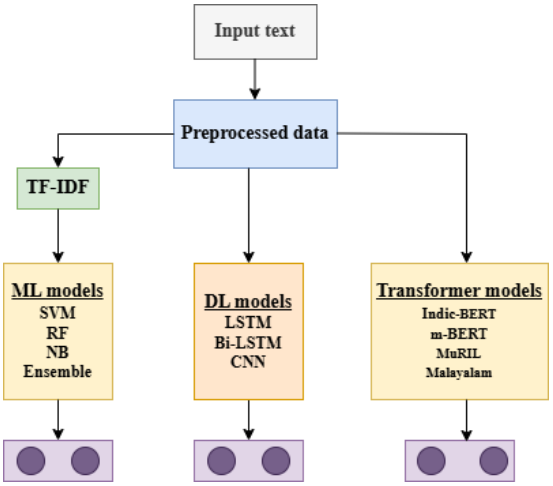

Figure 1: Methodological outline

4.1 Data preprocessing

As the dataset contains code-mixed text, proper preprocessing was crucial for ensuring meaningful feature representation. We have conducted several steps to achieve this. At first, we performed text cleaning was performed by removing emoticons, pictographs, URLs, and stopwords to eliminate noise. Next, to handle code-mixed text,

we used the Indic-Transliteration library to convert English words to Malayalam, ensuring consistency in linguistic representation. Finally, we employed subword-based tokenization for deep learning and transformer models to effectively process the mixed-language text.

4.2 Machine Learning Approaches

For the classification of the model, we used logistic regression (LR), support vector machine (SVM), random forest (RF), decision tree (DT) and naive bayes (NB). We used scikit-learn[1] to tune each model: max iterations of 1000 for LR, SVM with radial basis function(rbf) kernel, 1000 estimators for RF, unlimited depth for DT, and an alpha of 0.15 for NB. Majority voting with the ensemble model combined the LR, SVM, and DT models to increase robustness. The features were extracted using TF-IDF.

4.3 Deep Learning Approaches

We perform fake news classification using the LSTM, BiLSTM, and CNN models. The model development was done on the 'TensorFlow' framework. The text data has been tokenized and then converted into a padded sequence with a vocabulary size of 10,000 and a sequence length of 100. The LSTM model consists of two LSTM layers with 128 and 64 units, respectively. Each is followed by a 0.3 dropout layer to handle overfitting. The BiLSTM model consisted of two bidirectional LSTM layers of 128 and 64 units, respectively, using the same dropout strategy. The CNN model consisted of two 1D convolutional layers with 128 and 64 filters, kernel size 5, and max-pooling layers to capture the feature set. All of them used the same embedding layer of dimension 128, the Adam optimizer, and binary cross-entropy loss. The results were best after training up to a certain 10 epochs with a batch size of 32. Model performance was evaluated based on accuracy and F1 score.

4.4 Transformer Approaches

We fine-tuned pre-trained multilingual transformer models such as IndicBERT(Deode et al., 2023), MuRIL (Khanuja et al., 2021), mBERT (Devlin et al., 2019), and Malayalam (Joshi, 2023) from the Hugging Face[2] transformers library for our fake news classification task. For compatibility with each model, AutoTokenizer was instantiated with a sequence length of 256 tokens. We have used a batch size of 16, a learning rate of $2e^{-5}$, and trained for 10 epochs. We employed the Adam optimizer for gradient descent and used cross-entropy loss. We've trained the model using the dataset, and we've been evaluating its performance using accuracy and the classification report to measure its effectiveness in classifying fake news. Among all the model Malayalam has shown superior performance than all the other models.

Hyperparameter	Value
Batch Size	16
Optimizer	Adam
Epochs	10
Dropout Rate	0.3
Learning Rate	$2e^{-5}$

Table 2: Hyperparameter tuning

5 Result Analysis

This section presents a comparative performance analysis of various experimental approaches. The efficiency of the models is primarily assessed based on the F1-score, while precision and recall are also considered in some cases. A summary of the precision (P), recall (R), and F1-score for each model on the test set is presented in Table 3. Table 3 illustrates that for fake news classification, Malayalam perform the best. Among ML model, SVM, NB & Ensemble(SVM + DT + Logistic Regression) shows the same result. Among DL models, BiLSTM & CNN shows almost same result outperforming LSTM. Among Transformer, MuRIL shows a superior performance than any other models.

Method	Classifier	P	R	F1
ML	SVM	0.75	0.75	0.75
	RF	0.72	0.72	0.72
	NB	0.75	0.75	0.75
	Ensemble	0.75	0.75	0.75
DL	LSTM	0.25	0.50	0.33
	BiLSTM	0.81	0.81	0.81
	CNN	0.82	0.82	0.82
Transformers	Indic-BERT	0.76	0.75	0.74
	m-BERT	0.84	0.84	0.84
	MuRIL	0.83	0.82	0.82
	Malayalam	0.86	0.86	**0.86**

Table 3: Comparative analysis of performance on test data. Here P, R & F represent precision, recall & F1 score, respectively.

[1] https://scikit-learn.org
[2] https://huggingface.co/

6 Error Analysis

The misclassification occurred due to multiple challenges in the model's interpretation. It struggled with indirect speech and quotes. It misidentified sentences with numerical references, assuming numbers indicate factual accuracy. It also failed

Text	Actual label	Predicted label
ഓഷോ രജനീഷ് പറഞ്ഞപോലെ എനിക്കപ്പോൾ തോന്നിയത് അങ്ങനെയാണ് ഇപ്പോൾ തോന്നുന്നത് ഇങ്ങനെയാണ് എന്തൊക്കെയോ ആവോ	Fake	Original
ഫസ്റ്റ് ചോദ്യേ ജനഘല് അന്ധു വിദ്ദേ ബുസിനെസ്സ് ചേയ്ഥലുമ് വിജയികുമ് ഇന് ഇന്ദിഅ	Fake	Original
വ്ഹ്യ് അല്ല് ചോഉന്തിഏസ് ച്ലുബ് തോഗേഥേർ ന് ബ്ലോച്ച് ചിന തോ ഥേ മക്ഷിമുമ് ഏക്ഷ്തേന്ദ് അന്ദ് രേനമേ ചോവിദ് ഥൻ ചിനേസേ വിരുസ് ഇന് അല്ല് സുച് ചസേസ് വ്ഹേന് സോമേ ഓനേ വോർക് തോ ദേസ്തോയ് ഓഥേർസ് ഥേ നതുരേ ഇന്തേരേനേ ന് ചോന്തൈന് സുച് നേഗതിവേ ഭോർചേ ഥത് ഇസ് ഥേ ഗ്രചേ ഓഫ് ഗോദ് ഥേ ത്രുഥ്	Fake	Original

Figure 2: Misclasssification of text

to detect complex structures and misleading tones, mistaking deceptive content for original. These issues highlight the model's difficulty in recognizing nuanced patterns of misinformation. From Figure 3 and Figure 4, it can be seen that the Malayalam-Bert model correctly identified 383 fake news samples as fake and 468 original news samples as original. On the other hand, the mBERT model correctly

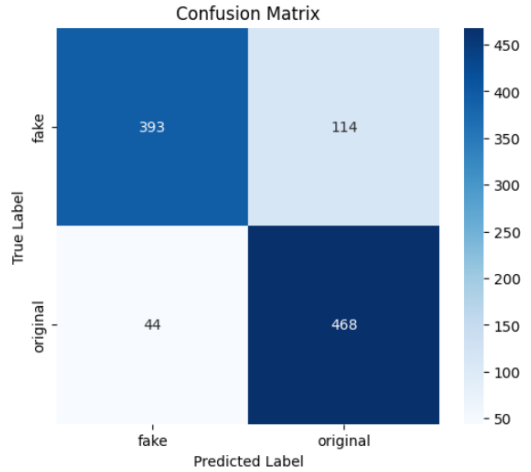

Figure 3: Confusion matrix of the Malayalam-BERT model

identified 427 fake news samples as fake and 428 original news samples as original. The confusion matrix of the top-performing

models (Malayalam and mBERT) is displayed in Figure 2, highlighting the highest precision achieved by the Malayalam model, as it correctly classifies most of the samples. Malayalam correctly identified 393 fake samples out of 507, whereas mBERT correctly identified 427 fake samples. Since mBERT successfully classifies more fake

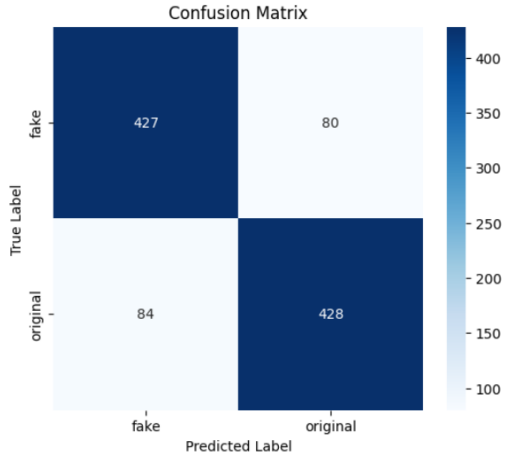

Figure 4: Confusion matrix of the m-BERT model

news samples, it can be concluded that mBERT performs better in fake news detection. However, the Malayalam model achieved the highest accuracy.

7 Conclusion

This paper discusses the detection of fake news in Dravidian languages by evaluating various ML, DL, and transformer-based approaches. Our experimental results document the best performance to be that of the Malayalam-BERT model with a maximum F1-score of 0.86 among the considered approaches. It further strengthens the efficiencies of transformer-based architectures which can handle complex linguistic structures in low-resource languages. In addition, future studies can be conducted by improving the performance of data augmentation, hybrid modeling techniques, and ensembling multiple transformer-based models to further improve robustness in fake news detection.

Limitations

While our approach demonstrates better performance, it has certain limitations also

- As Malayalam is a low-resourced language, it is difficult to capture its inherent linguistic complexities.
- Due to resource constraints, transformer ensembling could not be performed.

References

Samir Borgohain, Badal Soni, and Eduri Raja. 2023. Fake news detection in dravidian languages using

transfer learning with adaptive finetuning. *Engineering Applications of Artificial Intelligence*, 126.

Samruddhi Deode, Janhavi Gadre, Aditi Kajale, Ananya Joshi, and Raviraj Joshi. 2023. L3cube-indicsbert: A simple approach for learning cross-lingual sentence representations using multilingual bert.

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. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics.

Salman Farsi, Asrarul Eusha, Ariful Islam, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. CUET_Binary_Hackers@DravidianLangTech EACL2024: Fake news detection in Malayalam language leveraging fine-tuned MuRIL BERT. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, St. Julian's, Malta. Association for Computational Linguistics.

Raviraj Joshi. 2023. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *Preprint*, arXiv:2211.11418.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages. *Preprint*, arXiv:2103.10730.

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*, 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*, St. Julian's, Malta. Association for Computational Linguistics.

Eduri Raja, Badal Soni, and Sami Kumar Borgohain. 2024. Fake news detection in dravidian languages using multiscale residual cnn_bilstm hybrid model. *Expert Syst. Appl.*, 250:123967.

Eduri Raja, Badal Soni, and Samir Borgohain. 2023. *Fake News Detection in Dravidian Languages Using Transformer Models*, pages 515–523.

Kogilavani Shanmugavadivel, Malliga Subramanian, Sanjai R, Mohammed Sameer B, and Motheeswaran K. 2024. Beyond tech@DravidianLangTech2024 : Fake news detection in Dravidian languages using machine learning. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, St. Julian's, Malta. Association for Computational Linguistics.

G. Shimi et al. 2024. Language identification for dravidian languages: A crucial step for fake news detection in multilingual settings. *TBD*.

Symom Shohan, Md Hossain, Ashraful Paran, Shawly Ahsan, Jawad Hossain, and Moshiul Hoque. 2024. Semanticcuetsync at checkthat! 2024: Pre-trained transformer-based approach to detect check-worthy tweets.

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.

Xinyu Wang et al. 2024. A survey on monolingual and multilingual misinformation detection for low-resource languages. *TBD*.

Mesay Gemeda Yigezu et al. 2024. Ethio-fake: Integrating social-contextual and content-based features for fake news detection in under-resourced languages. *TBD*.