KCRL@DravidianLangTech 2025: Multi-Pooling Feature Fusion with XLM-RoBERTa for Malayalam Fake News Detection and Classification

Fariha Haq, Md. Tanvir Ahammed Shawon, Md. Ayon Mia, Golam Sarwar Md. Mursalin, Muhammad Ibrahim Khan
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u1904051, u1904077, u1804128}@student.cuet.ac.bd,
sarwarmursalin1015@gmail.com, muhammad_ikhan@cuet.ac.bd

Abstract

The rapid spread of misinformation on social media platforms necessitates robust detection mechanisms, particularly for languages with limited computational resources. This paper presents our system for the DravidianLangTech 2025 shared task on Fake News Detection in Malayalam YouTube comments, addressing both binary and multiclass classification challenges. We propose a Multi-Pooling Feature Fusion (MPFF) architecture that leverages [CLS] + Mean + Max pooling strategy with transformer models. Our system demonstrates strong performance across both tasks, achieving a macro-averaged F1 score of 0.874, ranking 6th in binary classification, and 0.628, securing 1st position in multiclass classification. Experimental results show that our MPFF approach with XLM-RoBERTa significantly outperforms traditional machine learning and deep learning baselines, particularly excelling in the more challenging multiclass scenario. These findings highlight the effectiveness of our methodology in capturing nuanced linguistic features for fake news detection in Malayalam, contributing to the advancement of automated verification systems for Dravidian languages.

1 Introduction

Social media platforms have turned out to be primary channels of information dissemination, generating massive volumes of data that require sophisticated techniques of analysis for verification of content authenticity (Farsi et al., 2024). The term "fake news" encompasses content spread without verification, published uncritically, and deliberately disseminated to cause social disorder (Majumdar et al., 2021). With the proliferation of social media platforms, manually verifying the authenticity of each piece of information has become increasingly challenging (Yigezu et al., 2024). The increasing impact of fake news on public opinion has highlighted the need for advanced detection systems, especially for low-resource languages. This work addresses the challenge of fake news detection in Malayalam YouTube comments through two distinct tasks: binary classification (Original vs. Fake) and multiclass classification into four categories ("FALSE", "HALF TRUE", "MOSTLY FALSE", and "PARTLY FALSE"). The task presents unique challenges due to Malayalam's linguistic complexity and the contextual nuances inherent in social media discourse. The following are the key contributions of this work:

- Development of a Multi-Pooling Feature Fusion (MPFF) architecture incorporating XLM-RoBERTa with CLS-Mean-Max pooling mechanism for robust Malayalam fake news detection
- Extensive experimentation on both binary and multiclass scenarios, establishing the efficacy of the model through rigorous performance metrics and comparative analysis.

The implementation details are publicly available in the GitHub repository:https://github.com/Ayon128/Shared-Task/tree/main/Fake%20News.

2 Related Works

Previous research demonstrates diverse approaches to fake news detection in Dravidian languages. (Rahman et al., 2024) achieved a 0.88 F1 score using a pre-trained Malayalam BERT model. (Tabassum et al., 2024) implemented XLMRoBERTa Base and BERT, reaching an F1 score of 0.87. (M et al., 2024) attained 0.86 using transformer models, while (Farsi et al., 2024) found that fine-tuned MuRIL BERT outperformed other multilingual BERT variants with an equivalent 0.86 F1 score. Transformer-based approaches by (Osama et al., 2024) using XLM-R and mBERT achieved 0.85, and (Tripty et al., 2024) reached 0.84 through

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 676–681

customized preprocessing with m-bert. For multiclass detection, (Kodali and Manukonda, 2024) explored BiLSTM classifiers with custom subword tokenizers, while (Anbalagan et al., 2024a) combined TF-IDF features with LaBSE embeddings in Naive Bayes models. (Anbalagan et al., 2024b) developed the MMFD framework achieving 86% accuracy using Gradient Boosting, and (Majumdar et al., 2021) implemented LSTM with word2vec embeddings, showing high training accuracy (98%) but lower validation accuracy (55%) due to overfitting. (Xing et al., 2024) comparatively analyze Mean, Max, and Weighted Sum pooling mechanisms for BERT and GPT in sentiment analysis, emphasizing task-dependent effectiveness.

3 Task and Dataset Description

The DravidianLangTech 2025 workshop presents a shared task on Fake News Detection in Dravidian Languages (Subramanian et al., 2025). This follows previous editions of similar tasks (Subramanian et al., 2023, 2024).The task is divided into two sub-tasks: Task 1 is binary classification to detect whether a text is "Original" or "Fake", and task 2 is multiclass classification to categorize text into four labels: "FALSE", "HALF TRUE", "MOSTLY FALSE", and "PARTLY FALSE". The datasets contain YouTube comments in the Malayalam language (Devika et al., 2024). As shown in Table 1, for task 1, the dataset consists of 5,091 texts divided into 3,257 texts for training, 815 texts for validation, and 1,019 texts for testing. Similarly, for task 2, the dataset consists of 2,100 texts divided into 1,900 texts for training and 200 texts for testing.

Task	Classes	Train	Dev	Test
Task A	Fake	1,599	406	507
	Original	1,658	409	512
Task B	FALSE	1386	-	100
	HALF TRUE	162	-	37
	MOSTLY FALSE	295	-	56
	PARTLY FALSE	57	-	7

Table 1: Distribution of Malayalam texts for binary (with dev set) and multi-class fake news detection tasks.

4 Methodology

We present an efficient framework for detecting fake news in Dravidian languages focusing on Malayalam text. Figure 2 illustrates an abstract overview of the entire system.

4.1 Preprocessing

We implement a systematic preprocessing pipeline to standardize the Dravidian Fake News dataset. The raw text is subjected to multiple cleaning processes, which includes removing URLs, handling emojis, eliminating hashtags and mentions, and normalizing sequential punctuation. This ensures consistent text representation before feeding into our models.

4.2 Augmentation

To address the severe class imbalance in Task 2, we applied random oversampling using scikit-learn. This was necessary due to the severe imbalance between the majority class (FALSE: 1386 samples) and minority classes, with a focus on "PARTLY FALSE", which has a mere 57 samples. We systematically oversampled minority classes ("HALF TRUE", "MOSTLY FALSE", and "PARTLY FALSE") with replacement to a total count of 200 samples in each category, thus greatly increasing representation across different categories. Without such a balanced sampling strategy, our model would suffer from strong bias towards the "FALSE" class, which forms about 66% of our training set, thus ignoring important subtleties in the severely underrepresented "PARTLY FALSE" category with just below 3% of the total samples. Our results show that such a sampling technique greatly improved minority class classification performance without negatively impacting general accuracy.

4.3 ML-based Approach

We employed several classical machine learning (ML) approaches for Dravidian fake news detection, including logistic regression (LR), support vector machine (SVM), random forest (RF), and XGBoost (XGB). Features were extracted from the preprocessed dataset using the TF-IDF vectorizer (Takenobu, 1994). The RF classifier was configured with 150 estimators and a minimum split threshold of 15 samples. For XGBoost, we utilized a learning rate of 0.05, 150 estimators, and a maximum depth of 6. The SVM classifier was implemented with a radial basis function (RBF) kernel, while Logistic Regression was configured with L2 regularization.

4.4 DL-based Approach

We implemented four deep learning architectures for Dravidian fake news detection, each processing

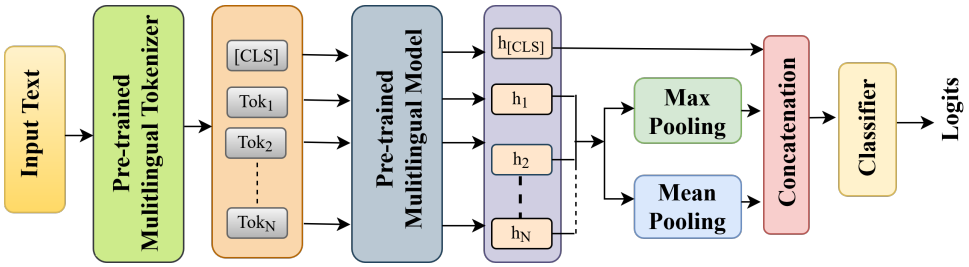

Figure 1: Architecture of the proposed Multi-Pooling Feature Fusion (MPFF) model utilizing XLM-RoBERTa for fake news detection, leveraging [CLS] token and mean and max pooling for enhanced classification.

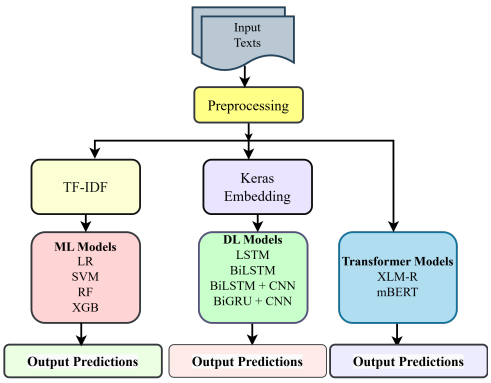

Figure 2: Overview of the Dravidian fake news detection pipeline incorporating traditional ML models with TF-IDF features, Deep Learning architectures using Keras embeddings, and Transformer-based approaches.

text through a 256-dimensional Keras embedding layer. The first architecture employs dual LSTM layers (64 and 32 units) with dropout (0.25) and batch normalization. The second utilizes a bidirectional LSTM with two layers (256 and 128 units). The third combines BiLSTM (256 units) with a dual-stage CNN (128 and 64 filters, kernel size 3), while the fourth substitutes GRU units for LSTM to optimize computational efficiency. We trained all models using Adam optimizer (learning rate 0.001) for 15 epochs with early stopping, implementing binary cross-entropy loss for Task 1 and categorical cross-entropy loss for Task 2.

4.5 Transformer-based Approach

For both tasks, we leveraged two multilingual transformer architectures: mBERT and XLM-RoBERTa (Devlin, 2018; Conneau, 2019), accessed via the Hugging Face platform (Wolf, 2019). The models were implemented in PyTorch with AdamW optimization, using batch size 32 and early stopping across 15 epochs. To ensure model robustness, we implemented k-fold cross-validation strategies: 10 and 7 folds for task 1 and task 2 respectively.

4.6 Multi-Pooling Feature Fusion (MPFF)

Our methodology introduces a Multi-Pooling Feature Fusion (MPFF) architecture, as shown in Figure 1, based on XLM-RoBERTa for both fake news detection tasks. The system processes input text through a pre-trained multilingual tokenizer, i.e., XLM-RoBERTa, generating tokens including a [CLS] token and sequence tokens. These are fed into the pre-trained multilingual model, producing hidden representations for each token. For both binary and multiclass classification tasks, our architecture implements parallel operations: [CLS] token extraction, max pooling, and mean pooling across token representations. These features are concatenated before passing through a classifier layer for final prediction output.

5 Experiments and Results

Table 2 presents the performance analysis across different model architectures. Our methodology introduces a Multi-Pooling Feature Fusion (MPFF) approach, implemented through [CLS] + Mean + Max pooling strategy in transformer architectures. For Task 1 (binary classification), traditional ML models demonstrated consistent performance, achieving macro-averaged F1 scores ranging from 0.735 to 0.775. Among DL approaches, BiLSTM achieved the best performance with F1 score 0.785, while BiGRU + CNN showed comparable results with F1 score 0.783, though BiLSTM + CNN significantly underperformed with F1 score 0.314. The MPFF strategy with XLM-RoBERTa significantly outperformed baseline approaches, achieving optimal results with F1 score 0.874 in Task 1, followed by mBERT with F1 score 0.862. For Task 2's multiclass scenario, our proposed MPFF approach with XLM-RoBERTa demonstrated superior performance with F1 score 0.628, significantly outperforming ML models and DL approaches. These results validate the effec-

Model	Pooling Strategy	Task 1 Performance			Task 2 Performance		
		Pr	Re	F1	Pr	Re	F1
ML Models							
LR	-	0.775	0.775	0.775	0.193	0.254	0.225
SVM	-	0.764	0.764	0.764	0.325	0.254	0.225
RF	-	0.743	0.735	0.735	0.442	0.294	0.294
XGB	-	0.773	0.752	0.752	0.315	0.283	0.283
DL Models							
LSTM	-	0.493	0.493	0.493	0.620	0.620	0.620
BiLSTM	-	0.785	0.785	0.785	0.505	0.505	0.505
BiLSTM + CNN	-	0.314	0.314	0.314	0.610	0.610	0.610
BiGRU + CNN	-	0.783	0.783	0.783	0.035	0.035	0.035
Transformer Models							
mBERT	[CLS]	0.858	0.861	0.860	0.710	0.588	0.608
	[CLS] + Mean + Max	0.866	0.862	0.862	**0.721**	0.592	0.622
XLM-RoBERTa	[CLS]	0.871	0.865	0.868	0.670	0.590	0.610
	[CLS] + Mean + Max	**0.875**	**0.874**	**0.874**	0.685	**0.597**	**0.628**

Table 2: Comparative performance analysis of model architectures on the test sets for Tasks 1 and 2. Pr, Re, and F1 denote macro-averaged precision, recall, and F1-score respectively.

tiveness of our MPFF approach in capturing comprehensive text representations for Malayalam fake news detection across both binary and multiclass scenarios.

6 Error Analysis

Analysis of the confusion matrices shown in Figures 3 and 4 reveals significant error patterns across classification tasks. For Task 1, despite achieving 82.8% accuracy in fake content detection (420 correct predictions), the model exhibits a systematic misclassification tendency where 17.2% of fake instances are incorrectly labeled as original. This error pattern predominantly occurs in texts containing subtle misinformation strategies that blend partial factual elements with deceptive content. In

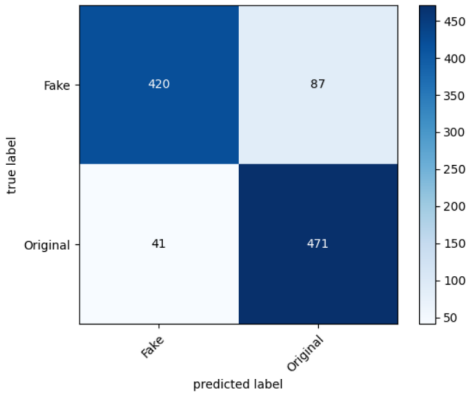

Figure 3: Confusion matrix showing the proposed model's binary classification performance for fake news detection.

the more challenging Task 2, our findings indicate a distinctive hierarchical confusion structure. The model demonstrates robust performance on the "FALSE" category (84% accuracy), yet shows progressive deterioration in performance across the spectrum of partial veracity categories. Specifically, "HALF TRUE" instances manifest confusion with both "FALSE" (12 instances) and "MOSTLY FALSE" (10 instances), while "PARTLY FALSE" classification achieves only 43% accuracy, constituting the model's most significant performance deficiency. Linguistic examination of misclassified samples reveals four primary error sources: contextual dependencies requiring domain-specific knowledge; subtle deceptive linguistic features in partially true content; interference between sentiment markers and factual assessment; and dialectal variations in informal discourse. These findings suggest that while current transformer architectures effectively capture binary veracity distinctions, they remain insufficiently calibrated to the nuanced spectrum of misinformation in low-resource languages like Malayalam.

7 Conclusions

In this study, we conducted a comprehensive analysis of fake news detection in Malayalam YouTube comments through binary and multiclass classification tasks. Our proposed Multi-Pooling Feature Fusion (MPFF) approach with XLM-RoBERTa achieved superior performance through effective integration of [CLS], Mean, and Max pooling features, obtaining macro-average F1 scores of 0.874 and 0.628 for binary and multiclass classification respectively.

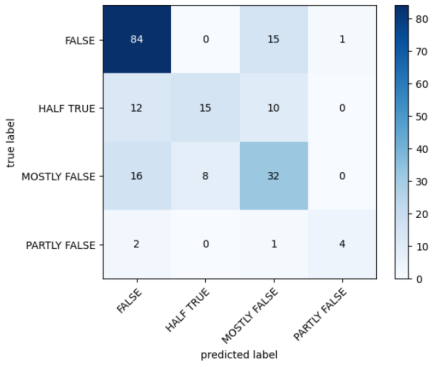

Figure 4: Confusion matrix showing the proposed model's multiclass classification performance for fake news categorization.

8 Limitations

Several limitations emerged in our work. First, despite using 5,091 texts for binary classification, our multiclass dataset of 2,100 texts remained insufficient and imbalanced. This limitation manifested in results, particularly in distinguishing nuanced categories. Second, our model demonstrated weakness in effectively classifying subtle forms of news content. Future work should focus on expanding the Malayalam datasets, exploring advanced architectures for improved classification performance, and investigating advanced Large Language Models (LLMs) which could potentially overcome the dataset limitations and better capture the nuanced distinctions between news categories that our current models struggled with.

References

Akshatha Anbalagan, Priyadharshini T, Niranjana A, Shreedevi Balaji, and Durairaj Thenmozhi. 2024a. WordWizards@DravidianLangTech 2024:fake news detection in Dravidian languages using cross-lingual sentence embeddings. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 162–166, St. Julian's, Malta. Association for Computational Linguistics.

Akshatha Anbalagan, Priyadharshini T, Niranjana A, Shreedevi Balaji, and Durairaj Thenmozhi. 2024b. WordWizards@DravidianLangTech 2024:fake news detection in Dravidian languages using cross-lingual sentence embeddings. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 162–166, St. Julian's, Malta. Association for Computational Linguistics.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv: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.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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*, pages 173–179, St. Julian's, Malta. Association for Computational Linguistics.

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.

Madhumitha M, Kunguma M, Tejashri J, and Jerin Mahibha C. 2024. TechWhiz@DravidianLangTech 2024: Fake news detection using deep learning models. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 200–204, St. Julian's, Malta. Association for Computational Linguistics.

Bhaskar Majumdar, Md. RafiuzzamanBhuiyan, Md. Arid Hasan, Md. Sanzidul Islam, and Sheak Rashed Haider Noori. 2021. Multi class fake news detection using lstm approach. In *2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART)*, pages 75–79.

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.

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.

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.

Tokunaga Takenobu. 1994. Text categorization based on weighted inverse document frequency. *Information Processing Society of Japan, SIGNL*, 94(100):33–40.

Zannatul Tripty, Md. Nafis, Antu Chowdhury, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. CUETSentimentSillies@DravidianLangTech EACL2024: Transformer-based approach for detecting and categorizing fake news in Malayalam language. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 245–251, St. Julian's, Malta. Association for Computational Linguistics.

T Wolf. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Jinming Xing, Ruilin Xing, and Yan Sun. 2024. Comparative analysis of pooling mechanisms in llms: A sentiment analysis perspective. *arXiv preprint arXiv:2411.14654*.

Mesay Yigezu, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2024. Habesha@DravidianLangTech 2024: Detecting fake news detection in Dravidian languages using deep learning. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 156–161, St. Julian's, Malta. Association for Computational Linguistics.