CUET_DUO@DravidianLangTech EACL2024: Fake News Classification Using Malayalam-BERT

Tanzim Rahman, Abu Bakkar Siddique Raihan, Md. Tanvir Rahman, Jawad Hossain, Shawly Ahsan, Avishek Das and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u1804015, u1804004, u1804002, u1704039, u1704057}@student.cuet.ac.bd
{avishek, moshiul_240}@cuet.ac.bd

Abstract

Identifying between fake and original news in social media demands vigilant procedures. This paper introduces the significant shared task on 'Fake News Detection in Dravidian Languages - DravidianLangTech@EACL 2024'. With a focus on the Malayalam language, this task is crucial in identifying social media posts as either fake or original news. The participating teams contribute immensely to this task through their varied strategies, employing methods ranging from conventional machine-learning techniques to advanced transformer-based models. Notably, the findings of this work highlight the effectiveness of the Malayalam-BERT model, demonstrating an impressive macro F1 score of **0.88** in distinguishing between fake and original news in Malayalam social media content, achieving a commendable rank of $\mathbf{1}^{st}$ among the participants.

1 Introduction

A growing number of people are choosing to get their news from social media platforms rather than traditional news outlets in an era where online interactions are becoming more common. News consumption on social media differs from traditional media, such as newspapers and television, regarding timeliness and affordability. Social media preference is also influenced by the ease with which news can be shared, commented on, and discussed with friends and other readers. However, the ease of sharing content via social media and the cost-effectiveness of online news distribution have led to the massive spread of fake news. This trend is particularly pertinent in the context of our paper on Fake News Detection in the Malayalam language. Kumari and Kumar (2021) introduced ensemble-based models for detecting offensive language in mixed-script social media posts. Many task formulations, datasets, and natural language processing (NLP) solutions have been used to investigate the intricacies of identifying fake news in the literature (Oshikawa et al., 2018).

This paper discusses the challenges of identifying fake news, focusing on the Malayalam language. The aim is to curb the spread of misinformation in this language space by providing culturally sensitive insights and solutions by exploring methodologies. Recent data on the impact of Facebook referrals, which shows that 20% of traffic goes to reliable websites and 50% goes to fake news sites, underscores the pressing nature of the problem. Considering that 62% of American adults get their news from social media, recognizing and mitigating the influence of fake content in online sources is more crucial than ever (Purcell et al., 2010).

This research contributes to the domain of fake news detection in the Malayalam language through the following key aspects:

- Investigate various machine learning, deep learning, and fine-tuned transformer models (m-BERT and Malayalam-BERT) to find the superior model for identifying fake news using relevant datasets.
- A comprehensive analysis of the proposed model to gain a nuanced understanding of its efficacy in recognizing fake news in Malayalam code-mixed social media content.

2 Related Work

Sivanaiah et al. (2022) prepared fake news datasets for several low-resource languages and applied Logistic Regression and BERT models for fake news classification. It was demonstrated through rigorous experiments that 'BERT-based-multilingual-cased' achieved a maximum F1 score of around 98%. At the same time, Logistic Regression reached approximately 95% in low-resource Indian languages such as Malayalam, Gujarati, and Tamil. Hariharan and Anand Kumar (2022) created a

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 223–228

multilingual low-resource fake news classification dataset and examined the impact of transformer-based models, such as multilingual BERT, XLM-RoBERTa, and MuRIL. For Telugu, Kannada, Tamil, and Malayalam, they assessed four transformer models: mBERT, XLM-RoBERTa, IndicBERT, and MuRIL. However, Raja et al. (2022) demonstrated that, for these low-resource languages, MuRIL had a higher accuracy in identifying fake news.

The DravidianLangTech@RANLP (Amjad et al., 2022) 2023 session "Fake News Detection in Dravidian Languages" concentrated on Malayalam content. In particular, the XLMRoBERTa-based model performed exceptionally well, obtaining a macro F1-score of 0.90. In DravidianLangTech-2023, Balaji et al. (2023) proposed transformer models such as M-BERT, ALBERT, BERT, and XLNET. M-BERT outperformed competitors with a robust F1 score of 0.74, surpassing XLNET and ALBERT, which achieved accuracy scores of 0.71 and 0.66, respectively. Using transformer-based models for language analysis, the study (Bala and Krishnamurthy, 2023) explored the nuances of identifying fake news. The *mural-base-cased* version of MuRIL was refined using a Dravidian language-curated dataset.

Rasel et al. (2022) addressed the scarcity of resources for the Bangla language in fake news detection by constructing a dataset of 4678 distinct news instances. Employing various machine learning, deep neural network, and transformer models, including CNN, CNN-LSTM, and BiLSTM, they achieved state-of-the-art accuracy ranging from 95.3% to 95.9%, showcasing notable improvements in accuracy and recall compared to previous studies when tested on both newly collected and existing datasets. Rahman et al. (2022) created the BFNC dataset containing 5,000 instances of fake news and presented the FaND-X framework using transformer-based and neural network-based techniques. With a maximum F1-score of 98% on the test data, experimental results showed that XLM-R outperformed other methods, demonstrating its efficacy in detecting fake news.

Abedalla et al. (2019) conducted a comparison of various BiLSTM models for detecting false information. This research demonstrates the efficacy of sequential models in text classification and identifying deceptive information, suggesting a potential future evaluation compared to BERT. In a similar application, the artificial intelligence initiative by Facebook (Kurasinski and Mihailescu, 2020) incorporates BERT as an integral component of its machine-learning strategy for detecting hate speech. In the past few years, pre-established models leveraging the Transformer architecture introduced (Vaswani et al., 2017) have gained prominence and serve a crucial function in sequence encoding and decoding. The effectiveness of these models in generating condensed contextualized embeddings for diverse texts inspired us to develop a system for detecting deceptive information based on these models.

3 Task and Dataset Descriptions

For the goal of fake news identification in the Malayalam language, the organizers created an almost balanced and standardized dataset. The primary purpose is to design a system that appropriately differentiates between fake and original news from social media posts in Malayalam. The dataset utilized in this challenge is derived from the corpus given by the workshop organizers (Subramanian et al., 2024). The work entails sorting social media statements into two predetermined classes: Fake and original news. Table 1 displays the distribution

Classes	Train	Test	Dev	TW
Original	1658	512	409	14031
Fake	1599	507	406	23198
Total	3257	1019	815	37229

Table 1: Distribution of Malayalam fake news dataset, where TW denote total words

of samples across the train, development (dev), and test sets for each class. Notably, the dataset is almost balanced, ensuring nearly equal instances for original and fake news classes.

4 Methodology

The proposed method experimented with several machine learning (ML), DL, and transformer-based baselines with fine-tuning the hyperparameters. Figure 1 demonstrates a schematic process of the employed models.

4.1 Feature Extraction

This work used TF-IDF (Sundaram et al., 2021) and Word2Vec embeddings (Rashid et al., 2020) for extracting textual features. The Keras embedding layer plays a crucial role in generating 100-

Method	Classifier	P	R	F1	A
ML	LR	0.92	0.26	0.41	0.62
	DT	0.84	0.31	0.45	0.63
	NB	0.78	0.79	0.78	0.78
DL	CNN	0.77	0.76	0.75	0.76
	BiLSTM	0.80	0.79	0.81	0.80
	CNN+BiLSTM	0.82	0.81	0.82	0.82
Transformers	m-BERT	0.73	0.52	0.38	0.52
	Malayalam-BERT	0.88	0.88	**0.88**	**0.88**

Table 2: Performance of various models for fake news classification in Malayalam, where P, R, F1, and A denotes precision, recall, macro F1-score, and accuracy, respectively

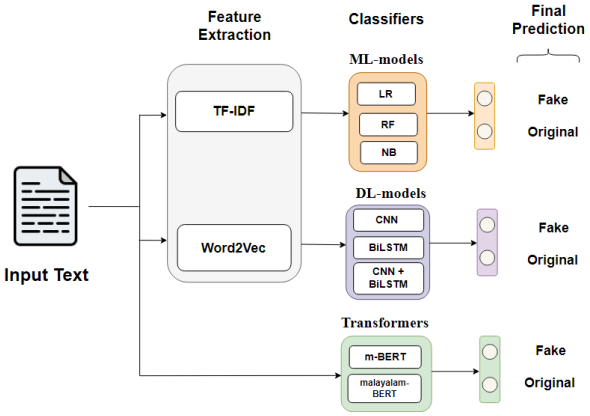

Figure 1: Schematic process of fake news identification

dimensional embedding vectors, enhancing the models' ability to identify and capture complex patterns in information, thereby improving the effectiveness of fake news detection.

4.2 ML Approaches

Various ML techniques, such as Logistic Regression (LR), Decision Tree (DT), and Naïve Bayes (NB), are explored for the task. This comprehensive approach involved meticulous parameterization to optimize the effectiveness of each algorithm. Specifically, the LR model underwent fine-tuning with a regularization value set at 0.01, while the DT is designed with a maximum depth of 10. Integrating NB included applying a radial basis function (RBF) kernel with a gamma value set to 0.001.

4.3 DL Approaches

A hybrid Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) architecture (Wu et al., 2020) was applied, featuring seven layers. Additionally, individual CNN and BiLSTM models were implemented as part of the broader hybrid architecture. The sequence vector of length 200 is input to the embedding layer, followed by two convolution layers with 'relu' activation and downsampling via max-pooling. The Bidirectional LSTM (BiLSTM) layer, with 128 units, captures intricate patterns, mitigating overfitting with a 0.5 dropout rate. The final layer uses a sigmoid activation function for binary classification, with variations exploring pre-trained word vectors. The 'Adam' optimizer employs a $1e^{-4}$ learning rate and binary cross-entropy as the loss function. Training spans 20 epochs with a batch size of 64, balancing performance and computational efficiency in fake news identification.

4.4 Transformer Models

This work applied two pre-trained transformer models, particularly M-BERT (Devlin et al., 2018), Malayalam-BERT (Joshi, 2022). These models, sourced from the Hugging Face transformers library [1] , underwent fine-tuning on the fake news corpus using the Ktrain (Maiya, 2022) package. The maximum sequence length was set at 100 with a batch size of 16. The models underwent training for three epochs, with a learning rate ($1e^{-4}$), enhancing their performance for the specific goal of fake news identification.

5 Results and Analysis

Table 2 illustrates the performance of the employed models for the fake news classification in Malayalam. Table 3 provides a comprehensive comparison of the performance across all participating teams. Our proposed model, Malayalam-BERT, has demonstrated superior performance, achieving the highest F1-score of **0.88** when compared to

[1]https://huggingface.co/

all other participating teams. illustrates a comparison of the performance of the opposing team with their respective ranks in the shared task. Among ML models, NB shines out with good accuracy of 0.78 and an impressive macro F1-score (0.78). On the other hand, CNN+BiLSTM achieved balanced accuracy (0.82), recall (0.81), and F1-score (0.82). Strategic modifications, including fine-

Team	F1_Score (Macro)	Rank
CUET_DUO	**0.88**	**1**
Punny_Punctuators	0.87	2
TechWhiz xlmr	0.86	3
CUET_Binary_Hackers	0.86	3
CUET_NLP_GoodFellows	0.85	4
CUETSentimentSilles	0.84	5

Table 3: Rank List of the Competition

tuning model hyperparameters and examining false positive occurrences, are proposed to increase the model's accuracy and overall efficacy in identifying fake news within the Malayalam language context.

6 Error Analysis

The fake news detection performance of the Malayalam-BERT model in Malayalam exhibits outstanding performance, especially evident in the high true positive count. Figure 2 shows the confusion matrix of the best-performed model (Malayalam-BERT) that highlighted the finest accuracy achieved by correctly tagging 442 out of 512 fake samples, demonstrating an effective capability to detect fake samples.

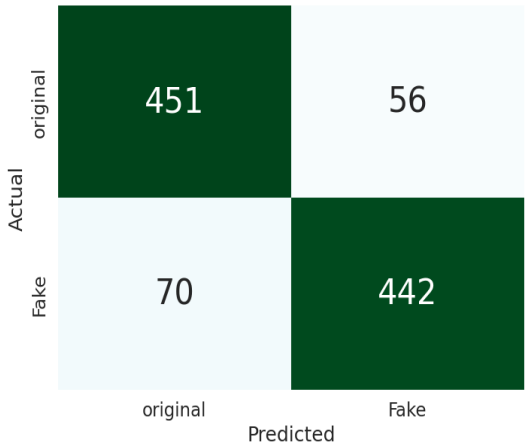

Figure 2: Confusion matrix of Malayalam-BERT model

However, a limitation arises in precision since 70 original samples were incorrectly identified as fake, indicating a vulnerability to false news detection. Less effective data preparation techniques could cause the misclassification. Addressing this issue through better preprocessing approaches may enhance precision and contribute to a more accurate classification. This observed pattern necessitates careful analysis and adjustment of the model's discriminatory capabilities. Further evaluations of the validation and test sets are essential to thoroughly examine the model's adaptability.

Figure 3 illustrates some predicted outcomes by the best-performed model (Malayalam-BERT).

Limitations

- The application of pre-trained transformers, specifically Malayalam-BERT, introduces a notable consideration in our fake news detection efforts. While leveraging pre-trained models can enhance contextual understanding, their effectiveness may be constrained by the specificity of the pre-training corpus. This may lead to a potential mismatch with the unique characteristics of fake news in Malayalam, highlighting the need for careful fine-tuning to ensure optimal performance.

Text Sample	Actual	Predicted
Sample1. കമ്മ്യൂണിസ്റ്റ് പൊളിഞ്ഞു തുടങ്ങി (The Communists began to collapse)	Original	Original
Sample2. ഇന്നെങ്കിലും ആ കണ്ണട മുഖത്ത് വെക്കും എന്ന് കരുതി പക്ഷേ വീണ്ടും ഞാൻ ശശിയായി (I thought I would put those glasses on my face at least today, but I was disappointed again)	Fake	Original
Sample3. ചൈന ഉത്പന്നങ്ങൾ ബഹിഷ്കരിക്കുക (Boycott China products)	Fake	Fake
Sample4. റാന്നിക്കാരെ മാത്രം കുറ്റം പറഞ്ഞവർ എന്തിയേ? ഇപ്പോൾ ചില യാ ളുകൾ കേരളം മുഴുവൻ പരത്തിയപ്പോൾ ആർക്കും കഴപ്പമില്ല (Why those who blamed only the Rannis? Now when some people have spread all over Kerala, no one has any problem)	Fake	Original
Sample5. കൊറോണ പോയി ഒന്ന് കുടെ മെച്ചപ്പെട്ട് ഓമൈക്രോനായി വന്നപ്പോൾ നമ്മുടെ പിണറവിന്റെ നേതൃത്വത്തിൽ ഒരു സ്വീകരണം കൊടുത്തതല്ലേ (When Corona went away and got better and came back as Omicron, didn't we give a reception under the leadership of our Pinu)	Original	Fake

Figure 3: Few examples of predicted outputs by the proposed (Malayalam-BERT) model

- Additionally, the inherent linguistic complexities of Malayalam pose challenges that may impact the model's ability to discern subtle patterns, warranting further investigation and refinement.

7 Conclusion

This work addresses the challenges of fake news detection in Malayalam by exploiting three ML, three DL, and two transformer-based models. Experimental investigations on the test dataset revealed

that Malayalam-BERT demonstrated superior performance among all models, achieving the highest macro F1 score (0.88). This finding highlights the proficiency of transformer-based strategies, precisely the efficiency of the Malayalam-BERT architecture, in excelling at the challenge of fake news identification. For future improvements, employing more language-specific preprocessing techniques and exploring ensemble models could enhance the overall performance of fake news detection in Malayalam. These strategies may contribute to refining the accuracy and robustness of the models in identifying misinformation effectively.

References

Ayat Abedalla, Aisha Al-Sadi, and Malak Abdullah. 2019. A closer look at fake news detection: A deep learning perspective. In *Proceedings of the 3rd International Conference on Advances in Artificial Intelligence*, pages 24–28.

Maaz Amjad, Sabur Butt, Hamza Imam Amjad, Alisa Zhila, Grigori Sidorov, and Alexander Gelbukh. 2022. Overview of the shared task on fake news detection in urdu at fire 2021. *arXiv preprint arXiv:2207.05133*.

Abhinaba Bala and Parameswari Krishnamurthy. 2023. Abhipaw@ dravidianlangtech: Fake news detection in dravidian languages using multilingual bert. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 235–238.

Varsha Balaji, B Bharathi, et al. 2023. Nlp_ssn_cse@ dravidianlangtech: Fake news detection in dravidian languages using transformer models. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 133–139.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

RamakrishnaIyer LekshmiAmmal Hariharan and Madasamy Anand Kumar. 2022. Impact of transformers on multilingual fake news detection for tamil and malayalam. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 196–208. Springer.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Jyoti Kumari and Abhinav Kumar. 2021. Offensive language identification on multilingual code mixing text. In *Working Notes of FIRE 2021-Forum for Information Retrieval Evaluation (Online). CEUR*.

Lukas Kurasinski and Radu-Casian Mihailescu. 2020. Towards machine learning explainability in text classification for fake news detection. In *2020 19th IEEE international conference on machine learning and applications (ICMLA)*, pages 775–781. IEEE.

Arun S Maiya. 2022. ktrain: A low-code library for augmented machine learning. *The Journal of Machine Learning Research*, 23(1):7070–7075.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2018. A survey on natural language processing for fake news detection. *arXiv preprint arXiv:1811.00770*.

Kristen Purcell, Lee Rainie, Amy Mitchell, Tom Rosenstiel, and Kenny Olmstead. 2010. Understanding the participatory news consumer. *Pew Internet and American Life Project*, 1:19–21.

MD Sijanur Rahman, Omar Sharif, Avishek Das, Sadia Afroze, and Mohammed Moshiul Hoque. 2022. Fand-x: Fake news detection using transformer-based multilingual masked language model. In *2022 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*, pages 153–158. IEEE.

Eduri Raja, Badal Soni, and Samir Kumar Borgohain. 2022. Fake news detection in dravidian languages using transformer models. In *International Conference on Computer Vision, High-Performance Computing, Smart Devices, and Networks*, pages 515–523. Springer.

Risul Islam Rasel, Anower Hossen Zihad, Nasrin Sultana, and Mohammed Moshiul Hoque. 2022. Bangla fake news detection using machine learning, deep learning and transformer models. In *2022 25th International Conference on Computer and Information Technology (ICCIT)*, pages 959–964. IEEE.

Umar Rashid, Muhammad Waseem Iqbal, Muhammad Akmal Skiandar, Muhammad Qasim Raiz, Muhammad Raza Naqvi, and Syed Khuram Shahzad. 2020. Emotion detection of contextual text using deep learning. In *2020 4th International symposium on multidisciplinary studies and innovative technologies (ISMSIT)*, pages 1–5. IEEE.

Rajalakshmi Sivanaiah, Nishaanth Ramanathan, Shajith Hameed, Rahul Rajagopalan, Angel Deborah Suseelan, and Mirnalinee Thanka Nadar Thanagathai. 2022. Fake news detection in low-resource languages. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 324–331. Springer.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, Premjith B, Sandhiya Raja, Vanaja, Mithunajha S, Devika K, Hariprasath S.B, Haripriya B, and Vigneshwar E. 2024. Overview of the Second Shared Task on Fake News Detection in Dravidian

Languages. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics.

Varun Sundaram, Saad Ahmed, Shaik Abdul Muqtadeer, and R Ravinder Reddy. 2021. Emotion analysis in text using tf-idf. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 292–297. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jheng-Long Wu, Yuanye He, Liang-Chih Yu, and K Robert Lai. 2020. Identifying emotion labels from psychiatric social texts using a bi-directional lstm-cnn model. *IEEE Access*, 8:66638–66646.