CUET_NLP_GoodFellows@DravidianLangTech EACL2024: A Transformer-Based Approach for Detecting Fake News in Dravidian Languages

Md Osama, Kawsar Ahmed, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u1804039, u1804017, u1804038, u1704039, u1704057}@student.cuet.ac.bd
moshiul_240@cuet.ac.bd

Abstract

In this modern era, many people have been using Facebook and Twitter, leading to increased information sharing and communication. However, a considerable amount of information on these platforms is misleading or intentionally crafted to deceive users, which is often termed as fake news. A shared task on fake news detection in Malayalam organized by DravidianLangTech@EACL 2024 allowed us for addressing the challenge of distinguishing between original and fake news content in the Malayalam language. Our approach involves creating an intelligent framework to categorize text as either fake or original. We experimented with various machine learning models, including Logistic Regression, Decision Tree, Random Forest, Multinomial Naive Bayes, SVM, and SGD, and various deep learning models, including CNN, BiLSTM, and BiLSTM + Attention. We also explored Indic-BERT, MuRIL, XLM-R, and m-BERT for transformer-based approaches. Notably, our most successful model, m-BERT, achieved a macro F1 score of 0.85 and ranked 4th in the shared task. This research contributes to combating misinformation on social media news, offering an effective solution to classify content accurately.

1 Introduction

In recent years, there has been an unprecedented surge in user participation on social media platforms such as Facebook and Twitter, as individuals increasingly utilize these platforms (Sharif et al., 2021). The users engage in the exchange of information, communication, and the continuous monitoring of current events. Conversely, a significant portion of the recent information disseminated on these platforms is inaccurately represented and, at times, deliberately crafted to misguide users. This content category is commonly identified as "fake news," encompassing any deceptive or false information presented as authentic news (Subramanian et al., 2024). The anonymity afforded to users on social media provides an opportunity for disseminators of fake news to manipulate people's beliefs, trust, and opinions by intentionally spreading false information. Rumors and misinformation propagate swiftly, adversely affecting personal relationships and social connections. Moreover, they have the potential to induce anxiety and emotional distress by fostering unfavorable perceptions, subjecting individuals to public scrutiny, and contributing to social isolation (Coelho et al., 2023). Moreover, current news often makes statements without confirmed evidence. To determine if these real-time claims are true, we heavily depend on how well they match information from other sources. The shared task (Subramanian et al., 2024) organized by DravidianLangTech@EACL 2024[1] provided us with an opportunity to address this significant challenge. This task aims to categorize a given social media text as either original or fake. The data sources include diverse social media platforms like Twitter and Facebook. The objective of this research work is to develop a system capable of discerning whether a news sample is original or fake. The key contributions of this endeavor are outlined below:

- Explored the efficacy of various ML, DL, and transformer models in detecting fake news and analyzing errors to gain valuable insights into the detection process.
- Proposed a transformer-based model that can classify a Malayalam news sample into two classes: fake and original.

2 Related work

The detection of fake news in low-resource languages, including code-mixed texts, is gaining increasing attention. Researchers have investigated

[1]https://sites.google.com/view/dravidianlangtech-2024/home

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 187–192

various techniques for identifying fake news using benchmarked corpora in low-resourced languages. In this section, we provide a concise overview of relevant studies in this domain. Coelho et al. (2023) addressed the challenge of detecting fake news through three machine learning models (MNB, LR, and Ensemble) trained on code-mixed Malayalam text using Term Frequency - Inverse Document Frequency (TF-IDF). They achieved a notable macro F1-score of 0.831 and secured 3^{rd} rank in the "Fake News Detection in Dravidian Languages" shared task at DravidianLangTech@RANLP 2023. In response to the urgent need for robust defenses against machine-generated fake news, Fung et al. (2021) created a benchmark dataset and identified fake news using cross-media consistency checking. Their proposed methodology surpassed the state-of-the-art models and achieved up to a 16.8% gain in accuracy. Rasel et al. (2022) constructed a comprehensive Bangla fake news dataset and have employed various machine learning, deep neural networks, and transformer models. The best performing models, CNN, CNN-LSTM, and BiLSTM, achieved notable accuracies of 95.9%, 95.5%, and 95.3%, respectively. Li et al. (2021) outlined the system for the AAAI 2021 shared task on COVID-19 fake news detection in English, securing the 3^{rd} position with a weighted score of 0.9859 on the test set. They constructed an ensemble of pre-trained language models, including BERT, Roberta, and Ernie, and employed diverse training strategies like a warm-up, a learning rate schedule, and k-fold cross-validation. Shu et al. (2019) addressed the challenge of fake news detection on social media by introducing the TriFN framework, a novel approach leveraging the inherent tri-relationship among publishers, news pieces, and users during dissemination. Unlike traditional algorithms focusing solely on news content, TriFN concurrently models publisher-news relations and user-news interactions. Zhou and Zafarani (2020) addressed the pressing issue of fake news, emphasizing its detrimental impact on democracy, justice, and public trust. Evaluating detection methods from multiple perspectives, including false knowledge, writing style, propagation patterns, and source credibility, the survey encourages interdisciplinary research. Sharif et al. (2021) presented a detailed description of a system developed for encompassing COVID-19 fake news detection in English (Task-A) and hostile post detection in Hindi (Task-B) using SVM, CNN, BiLSTM, and CNN+BiLSTM with TF-IDF and Word2Vec embedding. Their system achieved notable results, with the highest weighted F1 score of 94.39% in Task-A and 86.03% coarse-grained and 50.98% fine-grained F1 scores in Task-B.

3 Task and Dataset Description

The surge in online social media usage has revolutionized communication, enabling users to exchange information, engage in conversations, and stay informed about current events. However, this convenience has also led to the widespread dissemination of false information, commonly known as fake news, aiming to mislead users. This shared task (Subramanian et al., 2024) focuses on classifying social media texts as either original or fake news. The dataset comprises of text samples collected from diverse social media platforms, including Twitter and Facebook. It is organized into two distinct classes: "Fake" and "Original". The following outlines the definitions of the classes:

- **Fake:** Fake news refers to information deliberately crafted to mislead or deceive. These texts often contain intentionally false or misleading content that is presented as genuine.
- **Original:** Original news represents authentic and accurate information that reflects truthful and unbiased content. These texts are not manipulated or intentionally misleading, providing a reliable representation of real-world information.

Table 1 provides the distribution of samples in training, validation, and test sets across all the classes.

Classes	Train	Valid	Test
Fake	1,599	406	507
Original	1,658	409	512
Total	3,257	815	1,019

Table 1: Distribution of the dataset

4 Methodology

To address the issue at hand, we conducted an extensive exploration of various machine learning (ML), deep learning (DL), and transformer-based models. Through careful analysis, our research recommends utilizing a transformer-based model employing m-BERT (Jacob Devlin and Ming-Wei

Chang and Kenton Lee and Kristina Toutanova, 2018). Figure 1 provides a concise visualization of our methodology, outlining the key steps involved in our approach.

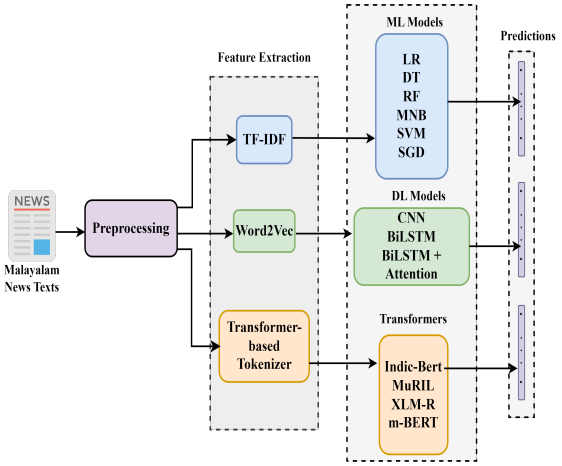

Figure 1: Abstract view of our methodology

4.1 Preprocessing

In the initial phase of our approach, we systematically executed essential preprocessing steps to refine the input data. This involved the meticulous removal of emojis, punctuation marks, URLs, and white spaces. By undertaking these measures, our objective was to optimize the quality and consistency of the dataset.

4.2 Feature Extraction

We employed a diverse set of techniques to capture and represent the underlying information within our textual data. The feature extraction techniques are as follows:

- **TF-IDF:** This technique (Qaiser and Ali, 2018) considers both the frequency of a term in a document and its rarity across the entire dataset, providing a robust representation of each document's content.
- **Word2Vec:** Leveraging the Word2Vec (Mikolov et al., 2013) technique, we transformed words into high-dimensional vectors, preserving semantic relationships and capturing context.
- **Transformer-based Tokenizer:** Leveraging transformer models, we used a transformer-based tokenizer to encode and tokenize our text data, benefiting from contextual information and hierarchical representations.

4.3 Model Building

In our research, we delved into a variety of models, including machine learning (ML), deep learning (DL), and transformer-based approaches.

4.3.1 ML models

In the realm of machine learning, our investigation involved the exploration and utilization of various classical models with TF-IDF. Specifically, we employed Logistic Regression, Decision Tree, Random Forest, Multinomial Naive Bayes, Support Vector Machine (SVM), and Stochastic Gradient Descent (SGD). Each of these models was strategically selected to harness different strengths and characteristics in addressing the complexities of this problem.

4.3.2 DL models

In this research work, we delved into the realm of deep learning models to get a better result with the word2vec word embedding technique. We experimented with a set of models, including CNN, BiLSTM (Huang et al., 2015), and BiLSTM + Attention (Vaswani et al., 2023), all incorporating word2vec embedding. Each model was chosen thoughtfully to extract unique insights and patterns from the data, contributing to a well-rounded analysis.

4.3.3 Transformer-based models

Finally, we delved into transformer-based models, specifically leveraging Indic-BERT (Jain et al., 2020), MuRIL (Khanuja et al., 2021), XLM-R (Conneau et al., 2019), and m-BERT (Jacob Devlin and Ming-Wei Chang and Kenton Lee and Kristina Toutanova, 2018) to enhance our outcomes. For these transformer models, we initially obtained them from the Hugging Face[2] library and fine-tuned them using our dataset. During testing, we streamlined the process using Hugging Face APIs, ultimately achieving accurate predictions.

5 Results

In this section, we provide comparisons of the performance achieved by different machine learning, deep learning, and transformer-based methods.

The performance evaluation of various classifiers for fake news detection showcases intriguing

[2]https://huggingface.co

Classifier	P	R	MF1
LR	0.67	0.65	0.64
DT	0.69	0.69	0.69
RF	0.71	0.71	0.71
MNB	0.72	0.71	0.71
SVM	0.70	0.69	0.69
SGD	0.73	0.73	0.73
CNN	0.78	0.72	0.71
BiLSTM	0.81	0.72	0.70
BiLSTM + Attention	0.80	0.72	0.71
Indic-BERT	0.67	0.64	0.66
MuRIL	0.74	0.76	0.75
XLM-R	0.84	0.83	0.84
m-BERT	**0.87**	**0.83**	**0.85**

Table 2: Performance of different models on test set

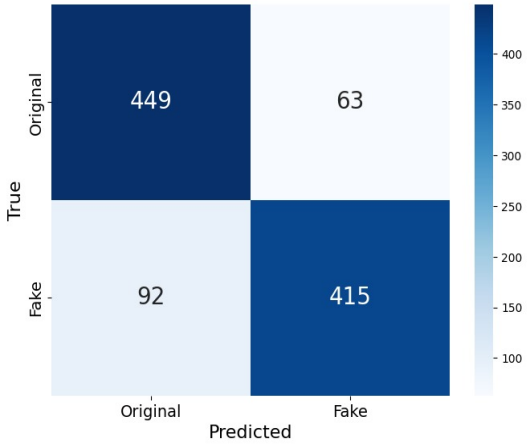

Figure 2: Confusion matrix of the m-BERT model for test set

insights into their predictive capabilities. A detailed summary of the precision (P), recall (R), and macro-F1 (MF1) scores attained by each model on the test set is provided in Table 2.

Among the traditional ML classifiers, Stochastic Gradient Descent (SGD) demonstrated the highest precision (P), recall (R), and macro-F1 (MF1) scores of 0.73, demonstrating consistent performance across all criteria.

Transitioning to deep learning architectures, CNN, BiLSTM, and BiLSTM + Attention exhibited competitive performances. While BiLSTM showed slightly higher precision (P), BiLSTM + Attention demonstrated better recall (R), underscoring the importance of attention mechanisms for discerning subtle patterns.

However, the standout performers were the transformer-based models, XLM-R and m-BERT. Outperforming other models in precision (P), recall (R), and macro-F1 (MF1) scores, m-BERT emerged as the top performer with the highest scores across all metrics, achieving a macro-F1 (MF1) score of 0.85.

5.1 Error Analysis

5.1.1 Quantitative Analysis

The results underscore the efficacy of transformer-based architectures, especially m-BERT, in detecting fake news. m-BERT's ability to leverage contextual information and encode multilingual text representations is pivotal in distinguishing between original and fake news samples.

This suggests that incorporating contextual embeddings from pre-trained language models, like m-BERT, significantly enhances the accuracy and robustness of fake news detection systems. Figure 2 represents the basic analysis with a confusion matrix. This matrix provides a quantitative breakdown of our model's predictions. The analysis reveals that our model accurately identifies 415 fake news articles and 449 original ones. However, there is a slight challenge in instances where the model misclassifies 92 fake news as original and 63 originals as fake. This occurrence is attributed to the presence of code-mixed text in certain samples of our dataset, leading to moments of confusion for the model.

5.1.2 Qualitative Analysis

Figure 3 showcases some sample predictions made by our model. Among these, samples 1, 2, and 4 are correctly classified. However, there are instances where the model misclassifies the samples, such as sample 3 being labeled as original when it

Text Sample	Actual	Predicted
Sample1: ഈ പാട്ടിന്റെ ആടിയ ചേച്ചിസിന്റെ തൊലിക്കട്ടി. (The skin of Chechis who swayed to this song.)	original	original
Sample2: താത്വിക ആചാര്യന്മാർക്ക് ഒരു കയ്യബദ്ധം... നാറ്റിക്കരുത് (A handout for philosophical teachers...don't be a jerk)	original	original
Sample3: ഇതൊക്കെ ഉള്ളത് തന്നെ.. (All this is there..)	Fake	original
Sample4: ചൈനയിലെ മരണ സംഖ്യ യൂറോ പ്പിനെക്കാൾ വലുതാണ് (China's death toll is higher than Europe's)	Fake	Fake
Sample5: ലോകമെന്നാൽ കണ്ണൂരും പരിസരപ്രദേശവുമായി ചുരുങ്ങിയോ? (Is the world reduced to Kannur and its surroundings?)	original	Fake

Figure 3: Some examples of predicted outputs by the best model. Here, corresponding English texts are translated using "Google Translator"

is actually fake, and sample 5 being inaccurately classified as fake being confused with code-mixed Malayalam texts. An imbalanced dataset might be the cause for this. Also, the use of code-mixed data in the corpus made it more difficult for the model to classify the text. These nuances highlight the importance of qualitative analysis in understanding the model's performance in specific cases.

Limitations

While our model achieved a commendable score in detecting fake news in Malayalam, certain limitations need consideration. These include the scarcity of diverse training data for Dravidian languages, potential linguistic nuances impacting model performance, and the model's focus primarily on textual content, neglecting multimedia elements often present in fake news. Additionally, the dynamic nature of misinformation tactics, the ethical implications of misclassification, cultural influences, and the need for explainability in model decisions pose ongoing challenges. Addressing these limitations will be crucial for refining the model's accuracy, adaptability, and ethical considerations in combating fake news effectively.

6 Conclusion

In our study, we set out to tackle the task of classifying fake and original news. Through a detailed comparison of various machine learning (ML), deep learning (DL), and transformer-based models, we found that m-BERT delivered the most impressive performance, boasting a macro F1 score of 0.85, surpassing all other models. Looking ahead, we plan to refine our approach further by exploring ensemble techniques in future research endeavors, aiming for an even more effective solution to combat misinformation.

References

Sharal Coelho, Asha Hegde, Kavya G, and Hosahalli Lakshmaiah Shashirekha. 2023. MUCS@DravidianLangTech2023: Malayalam Fake News Detection Using Machine Learning Approach. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 288–292, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avirup Sil. 2021. InfoSurgeon: Cross-Media Fine-grained Information Consistency Checking for Fake News Detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1683–1698.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging.

Jacob Devlin and Ming-Wei Chang and Kenton Lee and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805.

Kushal Jain, Adwait P. Deshpande, Kumar Shridhar, Felix Laumann, and Ayushman Dash. 2020. Indic-Transformers: An Analysis of Transformer Language Models for Indian Languages. *CoRR*, abs/2011.02323.

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 P. Talukdar. 2021. MuRIL: Multilingual Representations for Indian Languages. *CoRR*, abs/2103.10730.

Xiangyang Li, Yu Xia, Xiang Long, Zheng Li, and Sujian Li. 2021. Exploring Text-Transformers in AAAI 2021 Shared Task: COVID-19 Fake News Detection in English. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 106–115, Cham. Springer International Publishing.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Shahzad Qaiser and Ramsha Ali. 2018. Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications*, 181(1):25–29.

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.

Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2021. Combating hostility: Covid-19 fake news and hostile post detection in social media. *arXiv preprint arXiv:2101.03291*.

Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond News Contents: The Role of Social Context for Fake News Detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 312–320, New York, NY, USA. Association for Computing Machinery.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need.

Xinyi Zhou and Reza Zafarani. 2020. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Comput. Surv.*, 53(5).