CUET-NLP_Big_O@DravidianLangTech 2025: A BERT-based Approach to Detect Fake News from Malayalam Social Media Texts

Nazmus Sakib*, **Md. Refaj Hossan***, **Alamgir Hossain**
Jawad Hossain and Mohammed Moshiul Hoque
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
{u1904086, u1904007}@student.cuet.ac.bd, alamgir.hossain.cs@gmail.com
u1704039@student.cuet.ac.bd, moshiul_240@cuet.ac.bd

Abstract

The rapid growth of digital platforms and social media has significantly contributed to spreading fake news, posing serious societal challenges. While extensive research has been conducted on detecting fake news in high-resource languages (HRLs) such as English, relatively little attention has been given to low-resource languages (LRLs) like Malayalam due to insufficient data and computational tools. To address this challenge, the DravidianLangTech 2025 workshop organized a shared task on fake news detection in Dravidian languages. The task was divided into two sub-tasks, and our team participated in Task 1, which focused on classifying social media texts as original or fake. We explored a range of machine learning (ML) techniques, including Logistic Regression (LR), Multinomial Naïve Bayes (MNB), and Support Vector Machines (SVM), as well as deep learning (DL) models such as CNN, BiLSTM, and a hybrid CNN+BiLSTM. Additionally, this work examined several transformer-based models, including m-BERT, Indic-BERT, XLM-Roberta, and MuRIL-BERT, to exploit the task. Our team achieved 6th place in Task 1, with MuRIL-BERT delivering the best performance, achieving an F1 score of 0.874.

1 Introduction

In the digital era, social media platforms such as Facebook, Twitter, and Instagram have transformed how people share and consume information. These platforms let users stay updated on current events, express opinions, and participate in real-time global discussions. However, alongside these benefits, the rise of social media has also facilitated the proliferation of false or misleading information, commonly referred to as *fake news* (Subramanian et al., 2023). This phenomenon has become a critical concern due to its far-reaching consequences on public perception, societal trust, and decision-making processes. Fake news is content purposely created to misinform or deceive its audience, often impersonating reputable news sources (Subramanian et al., 2024). The rapid spread of fake news on social media exploits anonymity and platform reach, often outpacing factual content. The effects are severe, resulting in societal divisiveness, a loss of trust in credible news sources, and increased worry among individuals. Furthermore, bogus news can sway political decisions, harm reputations, and exacerbate existing societal divides (Farsi et al., 2024). Although significant progress has been achieved in detecting fake news in resourceful languages like English, less attention has been put towards low-resource languages, such as Malayalam, despite its speakers' rising digital footprint (Sharif et al., 2021). The lack of sufficient annotated datasets and the linguistic complexity of Malayalam pose unique challenges to building reliable fake news detection systems for this language. A shared task was organized under DravidianLangTech@NAACL 2025[1] to address this pressing issue, focusing on classifying social media texts into two categories: *Original* and *Fake* (Devika et al., 2024). As there is little research on Malayalam, we faced various difficulties like linguistic variations, dialect, and semantic identity (Coelho et al., 2023). The primary objective of this research is to design an efficient system capable of accurately classifying Malayalam news samples as fake or original, thus contributing to combating misinformation in low-resource languages. To achieve these objectives, our contributions to the task are as follows:

- Developed a transformer-based framework to detect fake news within the Malayalam dataset.

*Authors contributed equally to this work.

[1] https://sites.google.com/view/dravidianlangtech-2025/home

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 518–525
May 3, 2025 ©2025 Association for Computational Linguistics

- Investigated various ML, DL, and transformer-based models, evaluating their performance across metrics to identify the most effective model for detecting fake news in Malayalam. Presented an in-depth error analysis to refine the findings further.

2 Related Work

The proliferation of fake news on platforms like Facebook and Twitter often leads to misinformation and incorrect judgments. This growing concern has paved the way for research leveraging various ML and DL models in this domain (Sharif et al., 2021). While significant efforts have been made to address this issue, limited attention has been given to LRLs such as Malayalam. Different ML approaches have been devoted to a Malayalam dataset by Coelho et al. (2023) for fake news detection. They achieved the highest F1-score of 0.8310 using an ensemble of models (MNB+LR+SVM). In another study, M. San Ahmed (2021) developed a Kurdish dataset and applied ML models like LR, SVM, and Naive Bayes, with SVM achieving the highest accuracy of 88.17%. Additionally, Kumar and Singh (2022) employed ML models on a Hindi dataset containing 2,100 news articles to detect fake news, with Long Short-Term Memory (LSTM) achieving the highest F1-score of 0.89.

A recent study Krešňáková et al. (2019) utilized a fake news dataset from a competition and applied a CNN model, achieving an impressive F1-score of 0.97. Similarly, Kong et al. (2020) explored various neural network models on an English dataset, obtaining an accuracy of 90%. In another study, Kumar et al. (2020) developed a dataset by collecting data from Twitter, where a CNN+BiLSTM model with an attention mechanism achieved an accuracy of 88%. Additionally, Hiramath and Deshpande (2019) employed a dataset comprising news articles and found that a Deep Neural Network (DNN) achieved the highest accuracy of 91%. Several BERT variants, such as XLNet and ALBERT, outperformed deep learning approaches on a COVID-19 dataset (Gundapu and Mamidi, 2021). Schütz et al. (2021) utilized the *FakeNewsNet* dataset (Shu et al., 2020) and applied multiple transformer models, ultimately achieving the best F1 score of 0.84 with RoBERTa. Qazi et al. (2020) compared hybrid CNN models with transformer-based models, finding a slight improvement in F1 score to 0.47 with the transformer models. MuRiL-BERT also performed well on a Telugu dataset, achieving an F1 score of 0.87 (Hariharan et al., 2024). In another study, a comprehensive dataset for fake news detection in Bangla, a low-resource language, was developed, with LLM achieving the best F1 score of 0.89 (Shibu et al., 2025). A key limitation of past studies is their focus on HRLs, which results in biased models that may not transfer well to LRLs like Malayalam. In this context, we have presented a transformer-based framework tailored to handle Malayalam's unique linguistic and cultural aspects, improving detection accuracy for this underrepresented language.

3 Task and Dataset Description

The shared task[2] organizers provided a benchmark dataset for fake news detection in Malayalam (Subramanian et al., 2025). The dataset contains two classes: *Fake* and *Original*. The *Fake* class includes targeted texts, posts, or comments containing misinformation or falsified content, often created to mislead readers for political, commercial, or malicious purposes. The goal is to identify such content, which is especially common on social media during critical events. On the other hand, the class *Original* includes accurate, truthful posts providing reliable, verified information. The dataset contains 3,257 training samples, 815 development samples, and 1,019 test samples. Table 1 illustrated the class-wise distribution of the dataset.

Classes	Train	Dev	Test	W_T	UW_T
Original	1658	409	512	14031	8100
Fake	1599	406	507	23198	13100
Total	3257	815	1019	37229	19465

Table 1: Class-wise distribution of the dataset, where W_T and UW_T denote total words in three datasets and total unique words in train data.

The task's goal is to distinguish genuine news from fake news effectively. Figures A.1 and A.2 in Appedix A exhibit the word cloud distribution of classes.

4 Methodology

Several ML, DL, and transformer-based models are implemented and investigated to address the tasks. Figure 1 shows an outline of the methodology.

[2] https://codalab.lisn.upsaclay.fr/competitions/20698

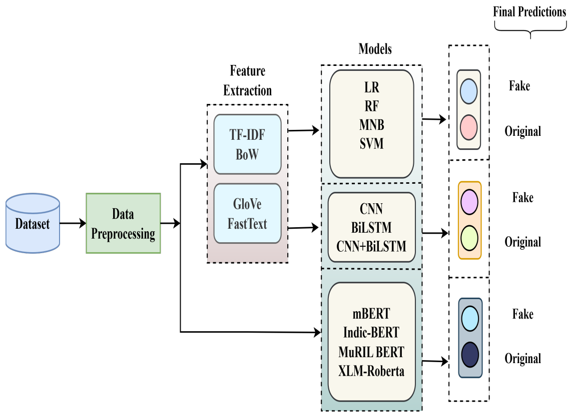

Figure 1: Schematic process of detecting fake news in Malayalam.

4.1 Data Preprocessing

Several preprocessing steps were applied to enhance the dataset's interpretability for the employed models. These steps included cleaning the text and removing unnecessary punctuation, emojis, and hyperlinks that could introduce noise into the data. Additionally, the MuRIL tokenizer was utilized to preprocess the text effectively. We used the MuRIL tokenizer with a maximum sequence length of 128 tokens.

4.2 Feature Extraction

For ML models, we utilized Term Frequency-Inverse Document Frequency (TF-IDF) and Bag of Words (BoW) representations with n-grams (unigrams and bigrams). We limited the vocabulary to the top 10,000 terms to balance interpretability and computational efficiency. We leveraged pre-trained word embeddings such as GloVe and FastText for DL models. Specifically, we used GloVe embeddings with a dimensionality of 120, which effectively captured word semantics and contextual relationships, and FastText embeddings trained on subword information to handle out-of-vocabulary words. We also employed MuRIL, a transformer-based model that tokenized the Malayalam text with a maximum token length of 128 and provided contextualized embeddings with 768 dimensions. These diverse embedding techniques ensured a robust text representation, enabling the models to accurately identify patterns and distinguish between fake and original news.

4.3 Classifiers

Four ML, six DL, and four transformer-based baselines are explored for fake news detection tasks.

4.3.1 ML Baselines

LR, SVM, RF, and MNB are utilized for the downstream task. The LIBLINEAR (Fan et al., 2008) solver function is used for ML models with GridSearchCV[3] to obtain better results. These traditional machine learning models serve as strong baselines to compare against transformer-based approaches, providing insight into the effectiveness of different learning paradigms. By leveraging GridSearchCV, we systematically tune hyperparameters to optimize each model's performance, ensuring a fair evaluation. This comparative analysis helps assess whether deep learning methods significantly outperform classical techniques in identifying misinformation.

4.3.2 DL Baselines

We employed CNN and the hybrid CNN+BiLSTM model for fake news detection, leveraging their ability to capture spatial and sequential patterns in textual data. The CNN model was designed to extract local features from the text using convolutional filters. Table 2 shows the fine-tuned hyperparameters for the deep learning-based models for the task.

Parameter	Value
Embedding Dimensions	128
Sequence Length	100
CNN Filters	64 filters of size 5
BiLSTM Units	64
Epochs	130
Batch Size	32
Optimizer	Adam
Learning Rate	1e-4

Table 2: Hyperparameter settings for CNN + BiLSTM model.

In contrast, the CNN+BiLSTM hybrid model combined the strengths of CNN's feature extraction with BiLSTM's ability to capture long-term dependencies and context. In our CNN+BiLSTM model, we configured a vocabulary size of 10,000, a sequence length of 100, and an embedding dimension of 128 for tokenization and embedding. The CNN branch was equipped with 64 filters of size 5 for local pattern extraction, and the BiLSTM branch had 64 units to capture bidirectional sequential relationships. The models were trained using

[3]https://scikit-learn.org/dev/modules/generated/sklearn.model_selection.GridSearchCV.html

sparse categorical cross-entropy as the loss function and the Adam optimizer with a learning rate of 1e-4. To address class imbalance, we computed class weights, ensuring fair model performance across categories.

4.3.3 Transformer-based Models

Transformer-based models were employed for fake news detection due to their ability to efficiently process large-scale contextual information, making them well-suited for multilingual tasks (Devlin et al., 2019). Several transformer models, including Indic-BERT (Dabre et al., 2022), mBERT (Pires et al., 2019), XLM-RoBERTa (Zhao and Tao, 2021), and MuRIL-BERT (Khanuja et al., 2021), were explored to evaluate their performance across diverse linguistic settings. Each model was fine-tuned for the classification task, with hyperparameters optimized to enhance performance. The MuRIL-BERT model demonstrated the best results, achieving an F1 score of 0.874. Table 3 presents the fine-tuned hyperparameters for the MuRIL-BERT model.

Parameter	Value
Batch Size	16
Epochs	9
Weight Decay	0.003
Learning Rate	2e-4

Table 3: Hyperparameter configuration for the transformer-based approach (MuRIL-BERT).

The model was trained using a learning rate of 2e-4, a weight decay of 0.003, and for 9 epochs. We trained on 9 epochs, as training for too many epochs (e.g., 10 or 15) led to overfitting, in which the model learns patterns too specific to the training data and loses generalization to unseen data. The optimal results highlight the effectiveness of MuRIL-BERT in handling the complex nature of fake news detection in multilingual datasets.

Additional implementation details can be accessed via the GitHub repository[4].

4.4 System Requirements

The model was trained on a dual GPU setup (NVIDIA Tesla T4x2), utilizing parallel processing for convolutional, BiLSTM, and transformer layers. The CNN+BiLSTM model required 5–8 GB of GPU memory and took approximately 60 minutes to complete training over 130 epochs. In contrast, the MuRIL-BERT model, which required 20 GB of GPU memory, completed training in just 20 minutes for 9 epochs. The training duration varied depending on the dataset size and the computation of class weights for handling class imbalances.

5 Result Analysis

Table 4 compares the performance of various classifiers for fake news detection, highlighting the precision (P), recall (R), F1 score, and G score.

Classifiers	Fake News Detection			
	P	**R**	**F1**	**G-Score**
LR	0.78	0.78	0.78	0.78
RF	0.78	0.77	0.77	0.77
MNB	0.80	0.80	0.80	0.80
SVM	0.80	0.79	0.79	0.79
CNN (F)	0.23	0.48	0.31	0.33
CNN (G)	0.24	0.48	0.31	0.34
BiLSTM (F)	0.28	0.51	0.36	0.38
BiLSTM (G)	0.27	0.51	0.35	0.37
CNN + BiLSTM (F)	0.29	0.49	0.36	0.38
CNN + BiLSTM (G)	0.29	0.48	0.36	0.37
Indic-BERT	0.81	0.82	0.81	0.81
m-BERT	0.83	0.81	0.82	0.82
XLM-R	0.86	0.85	0.86	0.85
MuRIL-BERT	**0.88**	**0.87**	**0.87**	**0.87**

Table 4: Performance of employed models on the test set, where F, G, and G-Score represent FastText, GloVe embeddings, and geometric mean score of precision and recall.

Among the ML models, MNB demonstrated an F1-score of 0.80, surpassing both LR (0.78) and SVM (0.79) in overall performance. This outcome indicates that MNB is better suited for this specific task, likely due to its efficiency in handling textual data distributions. Concerning DL models, CNN (G) and CNN (F) achieved F1 scores of 0.31, showing room for improvement in generalization. However, the hybrid CNN+BiLSTM models demonstrated a more robust performance. CNN+BiLSTM (F) achieved an F1 score of 0.36, outperforming both CNN (G) and CNN (F). This improvement highlights the strength of combining CNN's feature extraction capability with BiLSTM's sequential learning ability. However, CNN+BiLSTM (G) yielded a comparable performance with an F1-score of 0.36, slightly underperforming CNN+BiLSTM (F).

[4] https://github.com/Arghya-n/DravidianLangTech-FakeNews-2025

Transformer-based models significantly outperformed traditional and deep learning models because they efficiently process contextual information. Indic-BERT and m-BERT achieved F1-scores of 0.81 and 0.82, respectively, demonstrating strong performance for multilingual tasks. XLM-Roberta (XLM-R) further improved with an F1-score of 0.86, showcasing its capability in handling large-scale contextual information across diverse linguistic settings. Finally, MuRIL-BERT outperformed all other models, achieving the highest F1-score of 0.87 and G score of 0.87. The superior performance of MuRIL-BERT can be attributed to its robust contextual understanding, fine-tuned hyperparameters, and optimal training over nine epochs. This analysis highlights the consistent superiority of transformer-based models, particularly MuRIL-BERT, underscoring their ability to generalize well to multilingual and complex datasets. Appendix B presents a detailed error analysis of the proposed model's performance in detecting fake news in Malayalam. MuRIL-BERT performed well in detecting fake news from Malayalam social media texts due to its multilingual pretraining with a strong focus on Indian languages, including Malayalam. Unlike general multilingual models, MuRIL is trained on monolingual and transliterated text, allowing it to capture language-specific patterns common in social media. Fine-tuning domain-specific fake news data further enhanced its contextual understanding, enabling it to differentiate between misinformation cues, sentiment shifts, and propaganda techniques. This combination of pretraining advantages, contextual awareness, and careful optimization contributed to MuRIL-BERT achieving the best results in our experiments.

6 Conclusion

This work addressed the shared task by exploring various ML, DL, and transformer-based baselines for fake news detection in Malayalam. The results demonstrated that transformer-based models significantly outperformed others, with MuRIL-BERT achieving the highest F1-score of 0.87, demonstrating its superior capability to capture contextual information in multilingual datasets. Future work could explore advanced transformer architectures, such as GPT or ELMo, and integrate contextualized embeddings to enhance performance. Additionally, ensemble approaches that combine multiple transformer models or hybrid architectures tailored for fake news detection could offer even better results by leveraging the strengths of diverse models.

Limitations

The current work on fake news detection has several drawbacks, influenced by the following factors:

- Since the dataset is limited, the model's generalization is not guaranteed.
- Despite leveraging transformer-based models for contextual understanding, the system still struggles with detecting nuanced misinformation, such as subtle propaganda, satire, or region-specific deceptive narratives.

Acknowledgments

We thank the DravidianLangTech 2025 shared task organizers for running this task. This work was supported by the Directorate of Research & Extension (DRE), Chittagong University of Engineering & Technology (CUET).

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.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9(61):1871–1874.

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.

Sunil Gundapu and Radhika Mamidi. 2021. Transformer based automatic covid-19 fake news detection system. *Preprint*, arXiv:2101.00180.

R L Hariharan, Mahendranath Jinkathoti, P Sai Prasanna Kumar, and M Anand Kumar. 2024. Fake news detection in telugu language using transformers models. In *2024 5th International Conference on Innovative Trends in Information Technology (ICITIIT)*, pages 1–6.

Chaitra K Hiramath and G. C Deshpande. 2019. Fake news detection using deep learning techniques. In *2019 1st International Conference on Advances in Information Technology (ICAIT)*, pages 411–415.

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.

Sheng How Kong, Li Mei Tan, Keng Hoon Gan, and Nur Hana Samsudin. 2020. Fake news detection using deep learning. In *2020 IEEE 10th Symposium on Computer Applications Industrial Electronics (ISCAIE)*, pages 102–107.

Viera Maslej Krešňáková, Martin Sarnovský, and Peter Butka. 2019. Deep learning methods for fake news detection. In *2019 IEEE 19th International Symposium on Computational Intelligence and Informatics and 7th IEEE International Conference on Recent Achievements in Mechatronics, Automation, Computer Sciences and Robotics (CINTI-MACRo)*, pages 000143–000148.

Sachin Kumar, Rohan Asthana, Shashwat Upadhyay, Nidhi Upreti, and Mohammad Akbar. 2020. Fake news detection using deep learning models: A novel approach. *Transactions on Emerging Telecommunications Technologies*, 31(2):e3767. E3767 ETT-19-0216.R1.

Sudhanshu Kumar and Thoudam Doren Singh. 2022. Fake news detection on hindi news dataset. *Global Transitions Proceedings*, 3(1):289–297. International Conference on Intelligent Engineering Approach(ICIEA-2022).

Rania M. San Ahmed. 2021. Fake news detection in low-resourced languages "kurdish language" using machine learning algorithms. 12:4219–4225.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Momina Qazi, Muhammad U.S. Khan, and Mazhar Ali. 2020. Detection of fake news using transformer model. In *2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, pages 1–6.

Mina Schütz, Alexander Schindler, Melanie Siegel, and Kawa Nazemi. 2021. Automatic fake news detection with pre-trained transformer models. In *Pattern Recognition. ICPR International Workshops and Challenges*, pages 627–641, Cham. Springer International Publishing.

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

Hrithik Majumdar Shibu, Shrestha Datta, Md. Sumon Miah, Nasrullah Sami, Mahruba Sharmin Chowdhury, and Md. Saiful Islam. 2025. From scarcity to capability: Empowering fake news detection in low-resource languages with llms. *Preprint*, arXiv:2501.09604.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8(3):171–188.

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.

Yingjia Zhao and Xin Tao. 2021. ZYJ123@DravidianLangTech-EACL2021: Offensive language identification based on XLM-RoBERTa with DPCNN. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 216–221, Kyiv. Association for Computational Linguistics.

A Dataset Visualization

Despite leveraging transformer-based models for contextual understanding, the system still struggles with detecting nuanced misinformation, such as subtle propaganda, satire, or region-specific deceptive narratives.

Figure A.1 represents the most common words in fake news in the training set, which could indicate sensational language. In contrast, Figure A.2 shows the prominent words in the original news in the training set, reflecting the typical vocabulary of factual reporting. This analysis generated word clouds to visualize the most frequent words in fake and original news articles. A maximum of 200 words were used for visualization in each word cloud, with word size proportional to frequency.

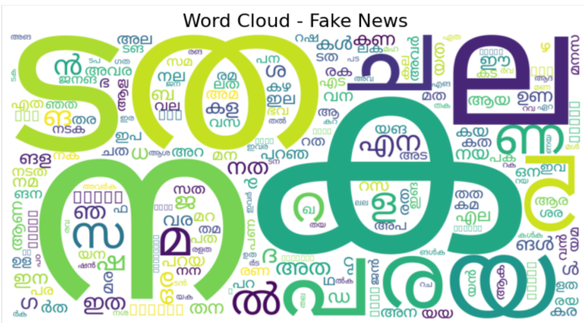

Figure A.1: Word Cloud distribution of *Fake* class.

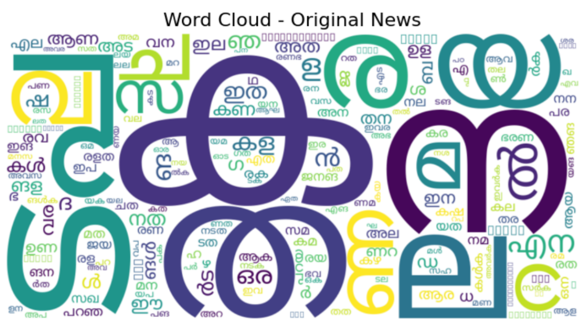

Figure A.2: Word Cloud distribution of *Original* class.

B Error Analysis

We have performed both quantitative and qualitative error analysis to obtain in-depth insights into the performance of the proposed model.

Quantitative Analysis: The MuRIL-BERT was used to conduct a quantitative error analysis, utilizing the confusion matrix shown in Figure B.1. The confusion matrix for the fake news detection

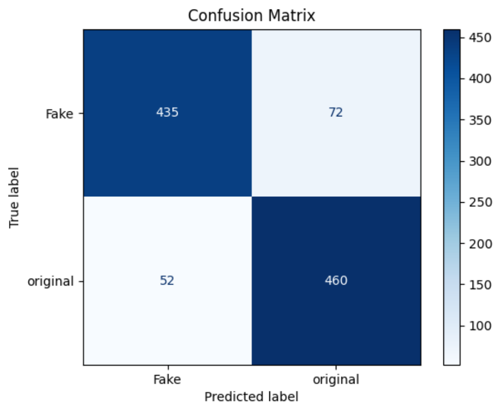

Figure B.1: Confusion matrix of the proposed model (MuRIL-BERT) for fake news detection.

task highlights the classification performance of the proposed MuRIL-BERT model. The model successfully classified most samples, with 435 instances of *Fake* and 460 instances of *Original* being correctly predicted. However, there were misclassifications: 72 *Fake* instances were misclassified as *Original*, while 52 *Original* instances were misclassified as *Fake*. These errors suggest that while the model performs well overall, challenges remain in distinguishing subtle differences between the *Fake* and *Original* categories, likely due to the dataset's overlapping linguistic patterns or contextual ambiguities. Further fine-tuning or incorporating additional contextual cues might improve the model's handling of such edge cases.

Qualitative Analysis: Figure B.2 depicts a qualitative analysis of the predictions made by the proposed MuRIL-BERT model for the fake news detection task. The model successfully classified samples 1 and 5 as *Fake* and samples 3 and 4 as *Original*, which aligns with their respective labels, showcasing its ability to identify a range of text samples correctly. However, the model incorrectly predicted sample 2 as *Original* instead of *Fake*, potentially due to linguistic nuances or overlapping features in the dataset.

Sample Text	Actual Label	Predicted Label
Sample-1: ചേട്ടാ വാർത്ത വയ്ക്കുന്നത് കേരളത്തിലാണ് സംഘി ഭരിക്കുന്ന നോർത്ത് ഇന്ത്യയിലല്ല,ഇവിടെ ആരോഗ്യ മന്ത്രി ഷൈലടീച്ചറാണ്	Fake	Fake
Sample-2: കൊറോണ സിപിഎം നേയും dyfi. യേയും ഭയക്കുന്നു. ബക്കറ്റിൽ പൈസയിടാൻ കാശില്ലാത്തതിനാൽ ഒളിച്ചോടാൻ തയാറെടുക്കുന്നു.	Fake	Original
Sample-3: തിരുവാതിര കളി നടക്കുമ്പോൾ ഗം ഓർത്തു ചിരിച്ചത് ഞാൻ മാത്രമാണോ?? 🤣	Original	Original
Sample-4: മന്ദബുദ്ധികളെ ഭരണഘടന സംരക്ഷിക്കുവാൻ ചുമതലപെടുത്തിയാൽ ഇങ്ങനെയൊക്കെ ഇരിക്കും	Original	Original
Sample-5: അവസരം നൽകൂ, ഏതെങ്കിലും വാദം ഉന്നയിക്കുമ്പോഴേക്കും ജയിലിൽ ഇടാൻ നോക്കുന്നതിനു എന്തിനാണ് , IMO പറയുന്നതിൽ വല്ല കാര്യവുമുണ്ടോ എന്നറിയേണ്ടേ	Fake	Fake

Figure B.2: Some predicted outputs by the proposed method (MuRIL-BERT).