CUET_Binary_Hackers@DravidianLangTech EACL2024: Fake News Detection in Malayalam Language Leveraging Fine-tuned MuRIL BERT

Salman Farsi, Asrarul Hoque Eusha, Ariful Islam, Hasan Mesbaul Ali Taher
Jawad Hossain, Shawly Ahsan, Avishek Das and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{salman.cuet.cse, asrar2860, arif.cse18cuet, Hasanmesbaul440}@gmail.com
{u1704039, u1704057}@student.cuet.ac.bd, {avishek, moshiul_240}@cuet.ac.bd

Abstract

Due to technological advancements, various methods have emerged for disseminating news to the masses. The pervasive reach of news, however, has given rise to a significant concern: the proliferation of fake news. In response to this challenge, a shared task in DravidianLangTech EACL2024 was initiated to detect fake news and classify its types in the Malayalam language. The shared task consisted of two sub-tasks. Task 1 focused on a binary classification problem, determining whether a piece of news is fake or not. Whereas task 2 delved into a multi-class classification problem, categorizing news into five distinct levels. Our approach involved the exploration of various machine learning (RF, SVM, XGBoost, Ensemble), deep learning (BiLSTM, CNN), and transformer-based models (MuRIL, IndicSBERT, m-BERT, XLM-R, Distil-BERT) by emphasizing parameter tuning to enhance overall model performance. As a result, we introduce a fine-tuned MuRIL model that leverages parameter tuning, achieving notable success with an F1-score of 0.86 in task 1 and 0.5191 in task 2. This successful implementation led to our system securing the 3^{rd} position in task 1 and the 1^{st} position in task 2. The source code will be found in the GitHub repository at this link: https://github.com/Salman1804102/DravidianLangTech-EACL-2024-FakeNews.

1 Introduction

Social media has gradually become an integral part of our lives, with regular posting and commenting being commonplace. Unfortunately, this platform is often misused, as individuals spread rumors by purposefully posting fake news to attack others and cause harm (Fowler, 2022; Medzerian, 2023). Given the importance of accurate information, it becomes crucial to curb the pervasiveness of fake news for the greater good. The widespread dissemination of false information carries catastrophic implications and potential dangers, particularly in the political and social spheres (De Paor and Heravi, 2020). A statistical analysis of the American public's aptitude for discerning between authentic and false news indicates a troubling pattern, as only 26% of survey participants express a high level of confidence in their ability to make this distinction (Watson, 2023). This low number underscores the urgent need for an automated system to detect fake news.

Numerous studies have been established for detecting fake news in high-resourced languages like English, Arabic, Spanish, French, German, etc. (Ahuja and Kumar, 2023; Mohawesh et al., 2023; Zhou et al., 2023; Al-Yahya et al., 2021; Guibon et al., 2019). But there is still much work to be done, especially for low-resourced languages like Malayalam, particularly in codemixed text (Thara and Poornachandran, 2021). Besides, in contrast to other Dravidian languages, Malayalam presents unique linguistic intricacies, encompassing dialect variations, nuanced word semantics, idiomatic expressions, and more (Coelho et al., 2023). These intricacies pose challenges in processing and analyzing Malayalam text. This shared task addresses precisely this issue, aiming to develop an automated system for detecting fake news and classifying its severity by categorizing news into various types. As participants in this shared task, the contributions of this paper are outlined as follows:

- We conducted a comprehensive comparative analysis of machine learning, deep learning, and transformer-based models through parameter tuning.
- We propose a fine-tuned MuRIL model that efficiently detects fake news, addresses class imbalance and performs news classification.

The rest of the paper follows this structure: Section 3 presents the task and dataset description, Section 4 discusses the methodology, Section 5

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 173–179

covers the results and error analysis, and Section 6 describes the conclusion and outlines future work.

2 Literature Review

The surge in fake news incidents prompted extensive research, initially relying on machine learning for detection. A specific study (Ahmed et al., 2017) harnessed machine learning and n-gram analysis, achieving a notable 92% accuracy on real news articles collected from Reuters. In the DravidianLangTech@RANLP 2023[1] shared task, the team 'MUCS' (Coelho et al., 2023) excelled with an impressive F1-score of 0.830. They employed ensemble models that combined Multinomial Naive Bayes (MNB), Logistic Regression (LR), and Support Vector Machine (SVM) to detect fake news in low-resourced Malayalam. As the dataset size expanded and the need for intricate detection across multiple languages arose, the forefront shifted towards employing deep learning methods. This is due to their efficiency in handling increased complexity. Incorporating four diverse datasets consisting of English-language news articles, this study (Sastrawan et al., 2022) utilized CNN, BiLSTM, and ResNet. It used popular word embeddings like Word2Vec, GloVe, and fastText. The results highlighted the consistent superiority of BiLSTM across all datasets. In a parallel investigation, Kumar and Singh (2022) addressed the proliferation of fake news in Hindi, drawing from diverse news articles. The author utilized NB, LR, and LSTM classifiers. With LSTM emerging as the most effective model for fake news detection, they achieved an impressive accuracy of 92.36%.

However, the shift from deep learning to transformer-based models like BERT has notably improved fake news detection accuracy, especially in code-mixed-text contexts (Kaliyar et al., 2021; Malliga et al., 2023). A study (Rahman et al., 2022) on a Bengali fake news dataset revealed XLM-R's superior accuracy of 98%. Whereas in the multilingual fake news classification scheme, Hariharan and Anand Kumar (2022) focused on low-resourced Tamil and Malayalam. They assessed the effectiveness of transformer-based models like m-BERT, XLM-R, and MuRIL. The focus on the low-resourced Dravidian languages has continuously increased. Sivanaiah et al. (2022) achieved an impressive F1-score of approximately 95% utilizing LR and 98% utilizing BERT models in one of the fake news detection endeavors for several Indian low-resourced languages like Tamil, Kannada, Gujarati, and Malayalam. The author (Thara and Poornachandran, 2021) in this study delved into the use of a dataset sourced from YouTube comments featuring Malayalam-English code-mixed text. The study explored the effectiveness of Camem-BERT, Distil-BERT, ELECTRA, and XLM-R models in this context. Remarkably, ELECTRA achieved an impressive F1-score of 99.33%. In another DravidianLangTech@RANLP 2023 study, Bala and Krishnamurthy (2023) implemented the MuRIL base variant model and achieved a notable F1-score of 87% for Malayalam code-mixed text. Within the context of fake news detection, addressing multi-class classification scenarios where news articles encompass varying degrees of truthfulness becomes crucial (Kaliyar et al., 2019; Karimi et al., 2018). A recent study (Shushkevich et al., 2023) has delved into this challenge, addressing the multi-class classification of fake news with labels such as 'True', 'Partially False', 'False', and others. To handle class imbalance, the researchers experimented with ChatGPT-based data augmentation, achieving an F1-score of 23% with m-BERT proving to be the most effective.

[1] https://dravidianlangtech.github.io/2023/

3 Task and Dataset Description

This shared task (Subramanian et al., 2024) on 'Fake News Detection in Dravidian Languages' consisted of two separate sub-tasks. In task 1, participants aimed to distinguish whether a post or comment is 'original' or 'fake'. Task 2 involved a more nuanced challenge, requiring participants to categorize news into five distinct labels: 'FALSE' (F), 'MOSTLY FALSE' (MF), 'PARTLY FALSE' (PF), 'MOSTLY TRUE' (MT), and 'HALF TRUE' (HT).

The competition organizers provided a dataset (Malliga et al., 2023) in multilingual and code-mixed Malayalam for these tasks. Task 1 comprised three distinct datasets (train, dev, and test), while task 2 involved two separate datasets (train and test). The training dataset in task 1 demonstrated a near balance, but task 2's training dataset exhibited a significant imbalance, with only a single occurrence of the MT class. The 'FALSE' class constituted about three-fourths of the samples. However, task 2 lacked a dedicated dev set. Further dataset statistics are detailed in Table 1.

Data	Class	Task 1			Class	Task 2		
		SC	UW	AL		SC	UW	AL
Train	Original	1,658	18,526	11	F	1,246	9,371	10
					MF	239		
					HT	141		
	Fake	1,599			PF	42		
					MT	1		
Dev	Original	409	5,581	11	Dev set is not present			
	Fake	406						
Test	Orignal	512	6,738	11	F	149	1,888	11
					MF	63		
					HT	24		
	Fake	507			PF	14		
					MT	0		

Table 1: Dataset statistics for both tasks, with acronyms SC, UW, and AL representing sample count, unique words, and average length, respectively.

4 Methodology

This section outlines the model framework devised to tackle the issue detailed in Section 2. Initially, to preprocess the text, we conducted several steps including the removal of emoticons, pictographs, URLs, and brackets. Following this, we employed feature extraction techniques to retrieve essential information. Our choice of feature extraction techniques was driven by the linguistic nuances inherent in such languages. TF-IDF (Takenobu, 1994) was selected for the ML models to effectively weigh term importance based on frequency, aligning with the need for interpretability in the context of fake news detection. On the other hand, GloVe embeddings (Pennington et al., 2014) were employed for the DL models, as they excel in capturing semantic relationships and contextual nuances within the text. This choice aimed to enhance the models' understanding of the intricacies present in Malayalam, contributing to more effective fake news classification. The overview of the proposed methodology is shown in figure 1.

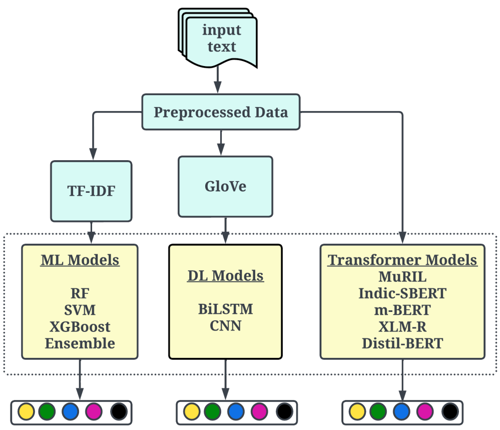

Figure 1: Overview of the methodology.

4.1 Machine Learning Approaches

In our exploration, we delved into several machine learning approaches, including Random Forest (RF), Support Vector Machine (SVM), eXtreme Gradient Boosting (XGBoost), and ensembles comprised of decision trees, SVM, and logistic regression. We used the TF-IDF feature extraction method for all ML models. To address the imbalanced dataset issue in Task 2, we assigned a class weight of type 'balanced' to all our ML approaches. The implementation of ML models was facilitated through the 'scikit-learn'[2] library for ease and efficiency. For our RF model, we set the 'n_estimators' to 100. With SVM, we utilized a 'linear' kernel with a regularization parameter C set to 1. Additionally, a tolerance of $1e^{-3}$ was employed as the stopping criterion. On the other hand, XGBoost was implemented with a 'multi:softmax' objective. For XGBoost, we used a learning rate of 0.3, 'n_estimators' of 100, and a maximum depth of 6. Lastly, an ensemble model consisting of LR, DT and SVM was utilized with a majority voting scheme.

4.2 Deep Learning Approaches

Deep learning models, BiLSTM and CNN, were implemented with a learning rate of $1e^{-3}$, 'Adam' as optimizer, and 'sparse_categorical_crossentropy' loss function. Class weights were employed to address the imbalanced data in task 2. The Bidirectional Long Short-Term Memory network (BiLSTM) was configured with a batch size of 64 and a single layer comprising 200 units. The Convolutional Neural Network (CNN) was designed with one layer containing 128 units and a batch size of 32. Both BiLSTM and CNN used the GloVe word embedding technique for feature extraction.

4.3 Transformer-based Approaches

We explored five transformer-based models, namely MuRIL (Khanuja et al., 2021), Indic-SBERT (Deode et al., 2023), m-BERT (Devlin et al., 2018), XLM-R (Conneau et al., 2020), and Distil-BERT (Sanh et al., 2019). All pre-trained transformer models were imported from 'Hugging Face' (Wolf et al., 2019) [3] and implemented using the ktrain library (Maiya, 2022). Subsequently, we fine-tuned these models on the provided datasets and utilized hyperparameter tuning to enhance

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

Method	Classifiers	Task 1			Task 2		
		P	R	F	P	R	F
ML models	RF (TF-IDF)	0.76	0.75	0.75	0.91	0.42	0.47
	SVM (TF-IDF)	0.76	0.76	0.76	0.58	0.44	0.45
	XGBoost (TF-IDF)	0.73	0.72	0.72	0.74	0.40	0.44
	Ensemble (TF-IDF)	0.77	0.77	0.77	0.51	0.29	0.31
DL models	BiLSTM (GloVe)	0.25	0.50	0.33	0.59	0.10	0.16
	CNN (GloVe)	0.31	0.46	0.29	0.43	0.18	0.17
Transformer models	**MuRIL**	0.86	0.86	**0.86**	0.66	0.48	**0.52**
	Indic-SBERT	0.86	0.86	0.86	0.33	0.24	0.17
	m-BERT	0.85	0.85	0.85	0.11	0.26	0.15
	XLM-R	0.86	0.86	0.86	0.12	0.25	0.16
	Distil-BERT	0.84	0.84	0.84	0.50	0.43	0.46

Table 2: Performance comparison of the proposed system over test data. Here P, R, and F stand for precision, recall, and macro F1-score respectively.

their performance. Specifically, in task 2, class weight augmentation was employed during the training of each transformer-based model to address class imbalance issues. We utilized the 'compute_class_weight' function imported from 'scikit-learn' for this purpose. The training configurations included a learning rate of $3e^{-5}$, batch size of 12 and 15 epochs for each respective model. For task 1 and task 2, we set the 'maxlen' parameter to 60 and 30, respectively. The rationale for choosing these parameters is grounded in a series of experiments conducted and GPU resource availability.

5 Results and Error Analysis

This section delves into the results and error analysis of the proposed fake news detection system. A detailed performance analysis is shown in Table 2.

5.1 Performance Analysis of Models

In task 1, among the different ML models, the Ensemble model outperformed others by achieving an F1-score of 0.77. Ensembling yielded improved performance by leveraging diverse strengths, enhancing generalization, and mitigating individual model weaknesses. Transformer-based models showcased superior performance in this task, and MuRIL turned out to be the best model by outperforming others. At the same time, the Indic-SBERT and XLM-R both displayed better results. As binary classification is inherently more straightforward for transformer-based models, this might contribute to the transformer models' effectiveness.

In task 2, ML models exhibited superior performance compared to transformer-based models except MuRIL. Mentionably, m-BERT's prediction skewed towards the 'HALF TRUE' class, and XLM-R consistently categorized maximum samples as 'FALSE' which led to extensive misclassification and poor performance. Since task 2 resembles a significant class imbalance, this issue also contributes to the differing performance of models. The ML models, being less complex, could potentially navigate the imbalanced dataset more effectively, resulting in superior performance compared to transformer-based models. Meanwhile, DL models encountered some specific challenges in both tasks. The poor performance of DL models in both tasks could be attributed to their sensitivity to the complexity and nuances of the Malayalam language. Apart from that, DL models, with their deep architectures, may overfit certain patterns in the training data, leading to a biased prediction tendency. Employing 'balanced' class weights in ML models was found to be better than using the 'compute_class_weight' function imported from 'scikit-learn' in the case of transformer-based models and DL models.

5.2 Error Analysis

Figure 2 shows the confusion matrix for task 1. It reveals that 433 fake news samples were correctly predicted, while 74 were misclassified. Similarly, 443 original news samples were accurately predicted, but 69 were erroneously classified as fake news in this task.

Moving to Figure 3, the confusion matrix for task 2 reveals insights into the model's performance on task 2. Among the 149 samples labeled as

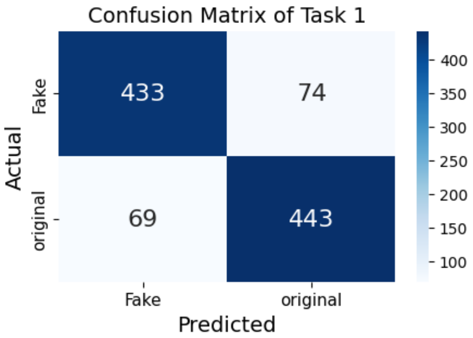

Figure 2: Confusion Matrix of task 1 for the MuRIL BERT.

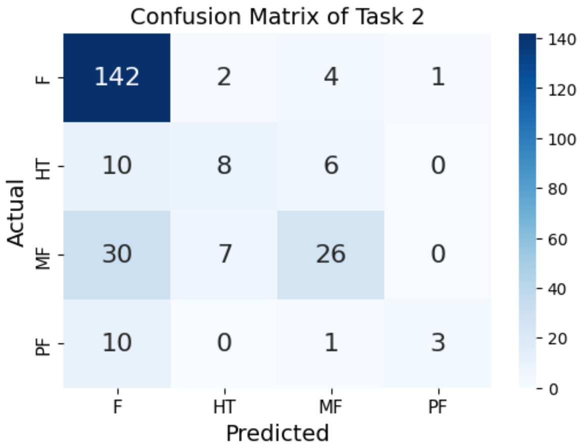

Figure 3: Confusion Matrix of task 2 for the MuRIL BERT.

'FALSE' in the test set, the model accurately predicted 142, with 7 misclassifications. However, for the 24 'HALF TRUE' samples, the model faced challenges, misclassifying 16 and achieving only 8 correct classifications. Among the 'MOSTLY FALSE' samples, 41.27% (26 out of 63) were accurately classified, while the remaining 58.73% faced misclassification. Notably, among the 'PARTLY FALSE' samples, 78.57% (11 out of 14) were misclassified. The elevated misclassification rates in these classes can be attributed to their limited number of instances in the dataset.

In summary, it appears the model predominantly predicted samples as 'FALSE', potentially influenced by the training data, where 'FALSE' samples comprised three-fourths of the dataset. A potential solution to address this issue could involve adjusting the efficient class weights mechanism, reducing the weight of the 'FALSE' class, and augmenting the weights of other classes for better model performance. However, some sample predictions for both tasks are provided in Appendix A.

5.3 Performance Comparison

Tables 3 and 4 show our position in the rank list.

Team Name	Score	Rank
CUET_DUO	0.88	1
Punny_Punctuators	0.87	2
CUET_Binary_Hackers	**0.86**	**3**

Table 3: A short rank list for task 1.

Team Name	Score	Rank
CUET_Binary_Hackers	**0.5191**	**1**
CUETSentimentSillies	0.4964	2
Quartlet	0.4797	3

Table 4: A short rank list for task 2.

Limitations

- The system is built on fine-tuning transformer-based models. It doesn't generalize to other languages and is not proven to give better results for a language that is not included in the training of MuRIL.
- Due to the GPU resource limitation, transformer ensembling couldn't be done.

6 Conclusion

This paper introduces a fake news detection system tailored for code-mixed Malayalam. It encompasses diverse models including ML, DL, and transformer-based models. Through extensive experimentation, evaluation, fine-tuning, and hyperparameter adjustments, the system showcases promising outcomes. Across both tasks, MuRIL emerges as the top performer, demonstrating its superior ability to handle code-mixed and transliterated Malayalam. The system achieves noteworthy F1-scores of 0.86 and 0.519, securing the 3^{rd} and 1^{st} positions in task 1 and task 2, respectively.

In the future, the exploration of fake news detection in other low-resourced Dravidian languages could be a worthwhile pursuit. Implementing data augmentation instead of relying solely on class weight adjustments for managing highly imbalanced datasets might prove more effective. Additionally, the utilization of hybrid models, combining transformers and DL models, holds the potential to yield improved results. Furthermore, exploring ensembles of transformer-based models could lead to superior performance.

References

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2017. Detection of online fake news using n-gram analysis and machine learning techniques. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1*, pages 127–138. Springer.

Nishtha Ahuja and Shailender Kumar. 2023. Mul-FaD: attention based detection of multiLingual fake news. *Journal of Ambient Intelligence and Humanized Computing*, 14(3):2481–2491.

Maha Al-Yahya, Hend Al-Khalifa, Heyam Al-Baity, Duaa AlSaeed, and Amr Essam. 2021. Arabic fake news detection: a comparative study of neural networks and transformer-based approaches. *Complexity*, 2021:1–10.

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.

Sharal Coelho, Asha Hegde, G Kavya, 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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Saoirse De Paor and Bahareh Heravi. 2020. Information literacy and fake news: How the field of librarianship can help combat the epidemic of fake news. *The Journal of Academic Librarianship*, 46(5):102218.

Samruddhi Deode, Janhavi Gadre, Aditi Kajale, Ananya Joshi, and Raviraj Joshi. 2023. L3Cube-IndicSBERT: A simple approach for learning cross-lingual sentence representations using multilingual BERT. *arXiv preprint arXiv:2304.11434*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gary Fowler. 2022. Council Post: Fake News, Its Impact And How Tech Can Combat Misinformation. Accessed on December 12, 2023.

Gaël Guibon, Liana Ermakova, Hosni Seffih, Anton Firsov, and Guillaume Le Noé-Bienvenu. 2019. Multilingual fake news detection with satire. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 392–402. Springer.

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.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2019. Multiclass fake news detection using ensemble machine learning. In *2019 IEEE 9th international conference on advanced computing (IACC), pages=103–107*. IEEE.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788.

Hamid Karimi, Proteek Roy, Sari Saba-Sadiya, and Jiliang Tang. 2018. Multi-Source Multi-Class Fake News Detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1546–1557, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. MuRIL: Multilingual representations for Indian languages. *arXiv preprint arXiv:2103.10730*.

Sudhanshu Kumar and Thoudam Doren Singh. 2022. Fake news detection on Hindi news dataset. *Global Transitions Proceedings*, 3(1):289–297.

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

S Malliga, Bharathi Raja Chakravarthi, SV Kogilavani, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, and Muskaan Singh. 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*, pages 59–63.

David Medzerian. 2023. Study Reveals Key Reason Why Fake News Spreads on Social Media. Accessed on December 12, 2023.

Rami Mohawesh, Sumbal Maqsood, and Qutaibah Althebyan. 2023. Multilingual deep learning framework for fake news detection using capsule neural network. *Journal of Intelligent Information Systems*, pages 1–17.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

I Kadek Sastrawan, IPA Bayupati, and Dewa Made Sri Arsa. 2022. Detection of fake news using deep learning CNN–RNN based methods. *ICT Express*, 8(3):396–408.

Elena Shushkevich, Mikhail Alexandrov, and John Cardiff. 2023. Improving Multiclass Classification of Fake News Using BERT-Based Models and ChatGPT-Augmented Data. *Inventions*, 8(5):112.

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.

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

S Thara and Prabaharan Poornachandran. 2021. Transformer based language identification for Malayalam-English code-mixed text. *IEEE Access*, 9:118837–118850.

Amy Watson. 2023. Confidence in Ability to Recognize Fake News in the U.S. Accessed on December 16, 2023.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Lina Zhou, Jie Tao, and Dongsong Zhang. 2023. Does fake news in different languages tell the same story? An analysis of multi-level thematic and emotional characteristics of news about COVID-19. *Information Systems Frontiers*, 25(2):493–512.

A Appendix

Figures 4 and 5 depict predictions corresponding to randomly selected samples from the dataset. The English translation of the Malayalam samples was generated using Google Translator.

Text Sample	Predicted	Actual
Sample 1: Ede sagavu endina epo America poyadu keraliti... (This is not going to continue, AP went to America, Keralite...)	original	fake
Sample 2: മോളെ. ഇത് കോമഡി സ്റ്റാർസ് അല്ല. ചിരിച്ചും കളിച... (Molly It's not comedy stars. Laugh and play...)	fake	original
Sample 3: Evar oke ntena verte virus nne indakan nadakn (Who knows when and where the virus will strike)	fake	original
Sample 4: Ethil appuram നാണക്കേഡ് വന്നിട്ടില്ല cpmne a... (After this, haven't seen the CPM around here, haha)	original	original
Sample 5: 2ദിവസം കൂടി കഴിഞ്ഞാൽ എല്ലാവർക്കും കൂടി വീട്ടിൽ... (If two days pass, everyone should stay at home together...)	original	original
Sample 6: ഇയാളെ കൊറോണ. രോഗി കൾ കിടയിൽ. ഇടാമായിരുന്നു---!! (He got Corona. Patients are increasing. Be cautious---!!)	fake	fake

Figure 4: Sample predictions for task 1 by MuRIL

Text Sample	Predicted	Actual
Sample 1: കണ്ണൂർ എയർപോർട്ടിന് ഭൂമി ഏറ്റെടുക്കുന്നതിന് എത... (What is the reason for acquiring land for Kannur Airport?)	Partly False	False
Sample 2: പലസ്തീൻ പതാകയണിഞ്ഞ് ക്രിസ്ത്യാനോ റൊണാൾഡോ. (Cristiano Ronaldo wearing the Palestinian flag.)	Half True	False
Sample 3: മലാപറമ്പ്-പുതുപ്പാടി റോഡ് നവീകരണം സംസ്ഥാന സർ.. (Malaparamp-Puthuppady Road Upgradation State Sir..)	Half True	Mostly Fales
Sample 4: പാലക്കാട് കൊലപാതകങ്ങളുടെ പശ്ചാത്തലിൽ ആഭ്യന്ത... (In the context of the Palakkad murders, domestic...	False	False
Sample 5: പാലക്കാട് ഇരുചക്രവാഹനങ്ങളിൽ പുരുഷന്മാരുടെ പിൻസ... (Men's rear in Palakkad two-wheeler...)	Half True	Half True
Sample 6: ബിവറേജസ് ഔട്ട്ലെറ്റുകൾ തുറന്ന ആദ്യ ദിനത്തില... (On the first day of opening of the Beverages outlets...)	Mostly False	Mostly False

Figure 5: Sample predictions for task 2 by MuRIL