One_by_zero@DravidianLangTech 2025: Fake News Detection in Malayalam Language Leveraging Transformer-based Approach

Dola Chakraborty[*], **Shamima Afroz**[*]
Jawad Hossain and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u1904012, u1904106, u1704039}@student.cuet.ac.bd, moshiul_240@cuet.ac.bd

Abstract

The rapid spread of misinformation in the digital era presents critical challenges for fake news detection, especially in low-resource languages like Malayalam, which lack the extensive datasets and pre-trained models available for widely spoken languages. This gap in resources makes it harder to build robust systems for combating misinformation, despite the significant societal and political consequences it can have. To address these challenges, we propose a transformer-based approach for Task 1 of the Fake News Detection in Dravidian Languages (DravidianLangTech@NAACL 2025), which focuses on classifying Malayalam social media texts as either *original* or *fake*. Our experiments involved a range of machine learning techniques, including Logistic Regression (LR), Support Vector Machines (SVM), and Decision Trees (DT), as well as deep learning architectures such as BiLSTM, BiLSTM-LSTM, and BiLSTM-CNN. Additionally, we explored transformer-based models, including IndicBERT, MuRIL, XLM-RoBERTa, and Malayalam BERT. Among these, Malayalam BERT achieved the best performance, with a macro F1-score of 0.892, securing us a rank of 3rd in the competition.

1 Introduction

Over the past several years, the proliferation of online social media has significantly transformed how individuals communicate, exchange information, and keep up with current affairs (Olan et al., 2024). Platforms such as Twitter, Facebook, and YouTube have enabled users to exchange information at an unprecedented scale (Sharif et al., 2021a). However, this convenience comes with a notable downside: a substantial portion of the information emerging on these platforms is false and, in many cases, intentionally designed to mislead users. Such content, commonly referred to as "fake news," encompasses any false or misleading information presented as original news (Melchior and Oliveira, 2024). The instantaneous reach of social media enables misinformation to spread rapidly, influencing public opinion and causing significant societal, organizational, and political repercussions. While considerable research has been conducted on fake news detection in high resource languages like Spanish, English (Sharma and Singh, 2024; Martínez-Gallego et al., 2021; Hu et al., 2024), low-resource Dravidian languages like Malayalam remain relatively underexplored. Malayalam, spoken mainly in the southern Indian state of Kerala (Thara and Poornachandran, 2022), presents unique linguistic challenges, such as dialect variations, limited annotated datasets and the complex morphology of the language. These aspects make fake news detection in Malayalam a particularly daunting task.

Existing attempts to address fake news in low-resource languages like Malayalam are often constrained by limited datasets, noisy code-mixed data, and the focus of state-of-the-art techniques on high-resource languages. To overcome these limitations, the shared task Fake News Detection in Dravidian Languages(Subramanian et al., 2024), organized by DravidianLangTech@NAACL 2025[1], introduces Task 1 which focuses on classifying social media texts and YouTube comments in Dravidian languages (Malayalam), as either *fake* or *original*.

As participants in this shared task, our work makes the following notable contributions:

- Proposed a transformer-based model specifically designed to classify Malayalam news content as "fake" and "original". This approach harnesses the capabilities of pre-trained language models to effectively tackle the challenges associated with processing

[*] Authors contributed equally to this work.

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

Malayalam, a low-resource language, in diverse domains such as social media posts and YouTube comments.

- Investigated a range of machine learning, deep learning, and fine-tuned transformer-based architectures to evaluate their effectiveness in detecting fake news and to analyze errors for deeper insights into the detection process.

The implementation of our proposed approach has been made publicly available, and the source code can be accessed on GitHub[2].

2 Related Work

Numerous studies have been conducted on fake news detection, primarily focusing on high-resource languages like English, while paying less attention to low-resource languages like Malayalam. Sharma and Singh (2024) pAhuja and Kumar (2023) proposed Mul-FaD, an attention-based model for fake news detection tested on English, German, and French news articles. The dataset, comprising 43,488 articles, was created by combining English datasets and translating parts into French and German. Mul-FaD achieved the best performance with 93.73% accuracy and an F1 score of 92.9, outperforming baseline models for multilingual fake news detection. Othman et al. (2024) explored Arabic fake news detection using hybrid models combining Arabic pre-trained BERT models (AraBERT, GigaBERT, MARBERT) with CNNs, with AraBERT-2D-CNN achieving the best F1-scores on Arabic datasets ANS (0.6188), Ara-News (0.7837), and Covid19Fakes (0.8009). Rahman et al. (2022) used the BFNC dataset for fake news detection, with XLM-R achieving the best performance, attaining an F1-score of 98% on the test data. Osama et al. (2024) performed the DravidianLangTech@EACL2024 shared task, tackling fake news detection in Malayalam using machine learning, deep learning, and transformer models. Their best model, m-BERT, achieved a macro F1-score of 0.85, ranking 4th and demonstrating its effectiveness in combating misinformation. Rahman et al. (2024) presented the shared task "Fake News Detection in Dravidian Languages - DravidianLangTech@EACL 2024," focusing on identifying fake and original news in Malayalam social media. Their teams employed diverse strategies, from machine learning to transformer models. Malayalam-BERT achieved the best performance with a macro F1-score of 0.88, securing 1st place. Farsi et al. (2024) conducted the DravidianLangTech@EACL2024 shared task focused on detecting fake news in Malayalam. Task 1 involved binary classification (fake or not), and Task 2 was multi-classclassification (five levels). Using machine learning, deep learning, and transformer models, they fine-tuned MuRIL, achieving F1-scores of 0.86 (Task 1) and 0.5191 (Task 2), securing 3rd place in Task 1 and 1st place in Task 2. Bala and Krishnamurthy (2023) implemented the MuRIL base variant model and achieved a notable F1-score of 87% for Malayalam code-mixed text. 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. Sharif et al. (2021b) presented a detailed description of a system developed for detecting COVID-19 fake news in English (Task-A) and hostile post detection in Hindi (Task-B) using SVM, CNN, BiLSTM, and CNN+BiLSTM with TF-IDF and Word2Vec embeddings. 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. Shyam and Poornachandran (2021) investigated a dataset of Malayalam-English code-mixed text from YouTube comments, evaluating models like Camem-BERT, Distil-BERT, ELECTRA, and XLM-R, with ELECTRA achieving an outstanding F1-score of 99.33%.

3 Task and Dataset Description:

This shared task (Subramanian et al., 2025) focuses on detecting fake news in the Dravidian language Malayalam. Task 1 requires classifying social media texts as either original or fake. The dataset (Devika et al., 2024; Subramanian et al., 2025, 2024, 2023), provided by the organizers, was curated from various social media platforms, including Twitter and Facebook. Table 1 illustrates the distribution of the dataset, which is fairly balanced. The training dataset consists of 1,658 original and 1,599 fake samples. Similarly, the validation dataset contains 409 original and 406 fake samples, while the test dataset includes 512 original and 507 fake sam-

[2] https://github.com/DolaChakraborty12/Fake_News_Detection_Dravidian_Language

ples.

Classes	**Train**	**Valid**	**Test**	$\mathbf{T_W}$
Original	1658	409	512	21626
Fake	1599	406	507	35629
Total	3257	815	1019	57255

Table 1: Dataset Statistics for Train, Validation, and Test Sets. (T_W denotes total words)

4 Methodology

We have implemented various ML, DL, and transformer-based approaches with hyperparameters fine-tuned to find out the best model for this task. Figure 1 depicts a schematic process in detecting fake news, illustrating each major phase.

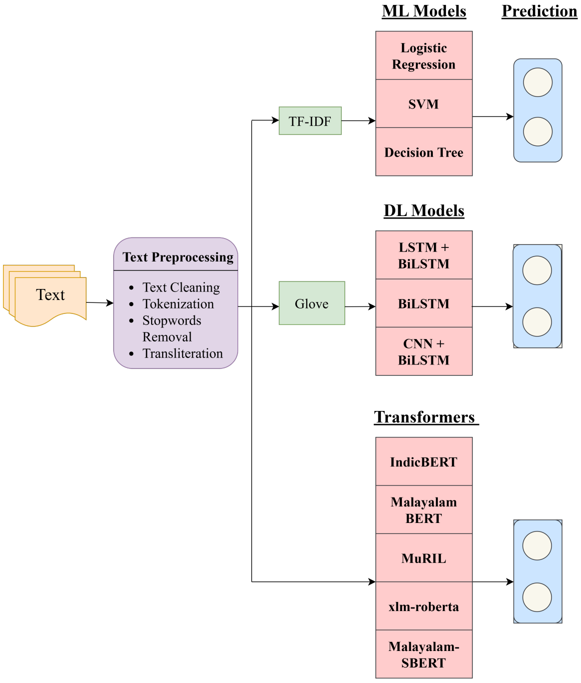


Figure 1: Schematic process of fake news detection.

4.1 Data Preprocessing

To ensure effective training and evaluation of our models, we have implemented an extensive data preprocessing pipeline. The pipeline ensured consistency and clarity in the datasets. The process involved removing punctuation, emojis, special characters, numerical text, and URLs to clean the data. HTML tags were eliminated to retain plain text, and frequent stopwords in both Malayalam and English were filtered out using the Malaya NLP library. Additionally, code-mixed text was standardized by transliterating it into Malayalam using AI4Bharat's transliteration engine. This streamlined preprocessing resulted in a refined dataset suitable for effective model training and evaluation.

4.2 Feature Extraction

Before passing text as input to machine learning, deep learning, and transformer-based models, it was first converted into a numerical format. For feature extraction, we used TF-IDF, GloVe, and Keras. TF-IDF assigns weights based on word frequency, with a vocabulary size of 10,000. GloVe provides pre-trained embeddings with an embedding matrix shape of (10,000, 100). The Keras embedding layer has an input dimension of 10,000 and an output dimension of 128, converting tokenized text into numerical sequences. This approach combines both statistical and semantic representations of text.

4.3 Machine Learning Models

For this task, we explored three machine learning models: Logistic Regression, Decision Trees, and Support Vector Machines (SVM). TF-IDF was employed for feature extraction. The Logistic Regression model utilized a regularization value of 0.01 to mitigate overfitting. The Decision Tree model was configured with a maximum depth of 10, while the SVM model employed a linear kernel for linear classification. Table 2 presents the hyperparameter of the machine learning based models.

Hyperparameter	**Value**
Max Depth (Decision Tree)	10
Regularization (Logistic Regression)	0.01
random state	42
Maximum Iterations	1000
class weight	balanced
Kernel (SVM)	linear

Table 2: Hyperparameters for machine learning models

4.4 Deep Learning Models

We utilized three deep learning models: BiLSTM, LSTM + BiLSTM, and CNN + BiLSTM, each starting with an embedding layer to process the input text.

- **BiLSTM Model:** This model employed a Bidirectional LSTM layer with 64 units to capture contextual patterns in both forward and backward directions. Dropout layers with rates of 0.8 and 0.5 were added to minimize overfitting.

- **LSTM + BiLSTM Model:** An additional LSTM layer with 64 units (returning sequences) was introduced before the BiLSTM layer. Dropout layers were included alongside an L2-regularized dense layer to enhance generalization. Dropout layer rates is 0.8 and 0.5.

- **CNN + BiLSTM Model:** This model combined a Conv1D layer with 128 filters and a kernel size of 5, followed by a MaxPooling1D layer for feature extraction, before passing the output to a BiLSTM layer with 64 units. To address overfitting, higher dropout rates of 0.9 and 0.8 were applied.

Table 3 presents the hyperparameters of deep learning based models.

Hyperparameter	Value
Optimizer	Adam
Learning Rate	2e-5
Epoch	10
Loss Function	Binary Cross-Entropy
Dropout (BiLSTM, LSTM+BiLSTM)	0.8, 0.5
Dropout (CNN + BiLSTM)	0.9, 0.8
Batch Size (BiLSTM, LSTM + BiLSTM)	16
Batch Size (CNN + BiLSTM)	32

Table 3: Hyperparameters for deep learning-based models

4.5 Transformer Models

Transformers have garnered significant attention in recent years due to their exceptional performance across various NLP tasks. For this task, we explored five pre-trained transformer-based models and fine-tuned them on our dataset to evaluate their effectiveness in this domain:

- MURIL: A multilingual model pre-trained on 17 Indian languages and English. It was fine-tuned with a batch size of 16, a learning rate of 1e-5, a sequence length of 60, and trained for 12 epochs. MURIL has proven to be highly efficient in multilingual tasks.

- IndicBERT: Pre-trained on 12 Indic languages along with English. It was fine-tuned with a batch size of 16, a learning rate of 2e-5, a sequence length of 60, and trained for 10 epochs.

- XLM-RoBERTa: Fine-tuned using the same hyperparameters as IndicBERT, but trained for 15 epochs.

- Malayalam-BERT: Specifically pre-trained on Malayalam text, tailored for tasks in the Malayalam language.

- Malayalam Sentence-BERT: Fine-tuned for sentence-pair tasks, optimized for Malayalam sentence-level tasks.

Table 4 presents the hyperparameters of the transformer models which are optimized through extensive experimentation.

Hyperparameter	Value
Optimizer	AdamW
Learning Rate	2e-5
Batch Size	16
Max Length	128
Epochs	15

Table 4: Hyperparameter setup for transformer-based models

5 Result Analysis

Table 5 illustrates the performance of the various ML, DL, and transformer-based models explored on the test dataset. The model's performance was evaluated using the macro F1-score. Transformer-based models, particularly Malayalam BERT, outperformed both ML and DL models, achieving the highest macro F1 score of 0.892. Among DL models, BiLSTM had the best score of 0.782, while the LR model led the ML models with a score of 0.5267. Overall, DL models performed better than ML models, but transformer-based models delivered superior performance overall.

5.1 Error Analysis

The performance of the best performed model is further investigated for in depth understanding of its behaviors using quantitative and qualitative error analysis.

5.1.1 Quantitative Analysis

Figure 2 presents the confusion matrix of the best-performing model, Malayalam BERT. A detailed quantitative error analysis of the fine-tuned Malayalam BERT model is performed based on the confusion matrix. It is evident from the confusion matrix that, out of 1,019 samples, 889 are correctly predicted. The model misclassifies 93 original samples as fake and 37 fake samples as original.

Model	P	R	F1
LR	0.75	0.75	0.75
SVM	0.76	0.76	0.76
DT	0.72	0.63	0.59
LSTM + BiLSTM(K)	0.80	0.80	0.80
BiLSTM(K)	0.80	0.80	0.80
CNN + BiLSTM(K)	0.81	0.81	0.81
LSTM + BiLSTM(G)	0.70	0.65	0.63
BiLSTM(G)	0.71	0.65	0.63
Malayalam BERT	**0.88**	**0.88**	**0.89**
IndicBERT	0.85	0.85	0.85
MuRIL	0.85	0.84	0.84
XLM-R	0.85	0.84	0.84
Malayalam S-BERT	0.86	0.86	0.86

Table 5: Performance of various ML, DL, Transformer-based models on the test set. P (Precision), R (Recall), F1 (macro F1-score)

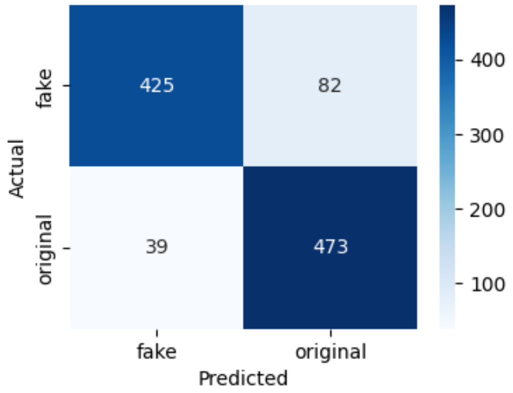
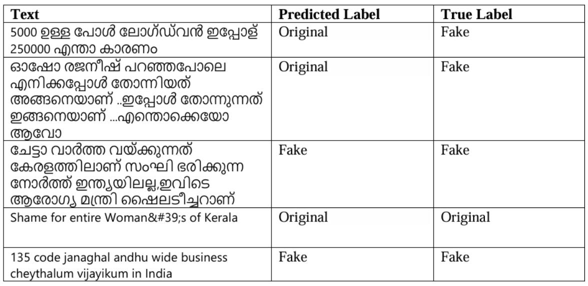


Figure 2: Confusion matrix of Malayalam BERT

5.1.2 Qualitative Analysis

A comparison of actual labels and predicted labels for a particular text is illustrated in Figure 3. The first two samples are incorrectly predicted as original, even though they are fake. However, the next three samples are predicted correctly as their actual classes.The misclassifications likely occur due to the linguistic complexity of Malayalam, including its rich morphology and syntactic structures, which pose challenges for the model in capturing subtle semantic differences. Although Malayalam BERT performs exceptionally well for the Malayalam language, fake news often imitates the style and tone of genuine news, making it challenging for the model to distinguish between the two.



Text	Predicted Label	True Label
5000 ഉള്ള പോൾ ലോഗ്ഡ്വൻ ഇപ്പോള് 250000 എന്താ കാരണം	Original	Fake
ഓഷോ രജനീഷ് പറഞ്ഞപോലെ എനിക്കപ്പോൾ തോന്നിയത് അങ്ങനെയാണ് ..ഇപ്പോൾ തോന്നുന്നത് ഇങ്ങനെയാണ് ...എന്തൊക്കെയോ ആവോ	Original	Fake
ചേട്ടാ വാർത്ത വയ്ക്കുന്നത് കേരളത്തിലാണ് സംഘി ഭരിക്കുന്ന നോർത്ത് ഇന്ത്യയിലല്ല,ഇവിടെ ആരോഗ്യ മന്ത്രി ഷൈലടീച്ചറാണ്	Fake	Fake
Shame for entire Woman&#39;s of Kerala	Original	Original
135 code janaghal andhu wide business cheythalum vijayikum in India	Fake	Fake

Figure 3: A few examples of predicted outputs by the Malayalam BERT

6 Conclusion

This paper evaluates the performance of various machine learning, deep learning, and transformer-based models for detecting fake news in Malayalam. While deep learning techniques such as LSTM + BiLSTM, BiLSTM, and CNN + BiLSTM demonstrated strong results, traditional machine learning methods struggled to effectively capture the intricate semantic relationships inherent in the Malayalam language. Among all approaches, Malayalam BERT achieved the best performance, with an F1-score of 0.892, by effectively capturing the language's unique nuances. Future research could focus on enhancing this work by utilizing larger datasets, leveraging ensemble transformer models, and exploring other advanced large language models.

Limitations

Our current work posses some limitations. Some limitations of our work are: i) Malayalam BERT has token limitation, causing truncation of long news articles and potential loss of crucial information. ii) Our task only analyzes text, making it ineffective against misinformation spread via images, memes, or misleading visuals. iii) Due to the small dataset size, the model may struggle to generalize well to diverse fake news patterns.

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

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.

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, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Varsha Balaji, Shahul Hameed T, and Bharathi B. 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, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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.

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.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22105–22113.

Kevin Martínez-Gallego, Andrés M Álvarez-Ortiz, and Julián D Arias-Londoño. 2021. Fake news detection in spanish using deep learning techniques. *arXiv preprint arXiv:2110.06461*.

Cristiane Melchior and Mírian Oliveira. 2024. A systematic literature review of the motivations to share fake news on social media platforms and how to fight them. *new media & society*, 26(2):1127–1150.

Femi Olan, Uchitha Jayawickrama, Emmanuel Ogiemwonyi Arakpogun, Jana Suklan, and Shaofeng Liu. 2024. Fake news on social media: the impact on society. *Information Systems Frontiers*, 26(2):443–458.

Md Osama, Kawsar Ahmed, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. CUET_NLP_GoodFellows@DravidianLangTech EACL2024: A transformer-based approach for detecting fake news in Dravidian languages. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 187–192, St. Julian's, Malta. Association for Computational Linguistics.

Nermin Abdelhakim Othman, Doaa S Elzanfaly, and Mostafa Mahmoud M Elhawary. 2024. Arabic fake news detection using deep learning. *IEEE Access*.

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.

Tanzim Rahman, Abu Raihan, Md. Rahman, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. CUET_DUO@DravidianLangTech EACL2024: Fake news classification using Malayalam-BERT. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 223–228, St. Julian's, Malta. Association for Computational Linguistics.

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

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

Upasna Sharma and Jaswinder Singh. 2024. A comprehensive overview of fake news detection on social networks. *Social Network Analysis and Mining*, 14(1):120.

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

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.

S Thara and Prabaharan Poornachandran. 2022. Social media text analytics of malayalam–english code-mixed using deep learning. *Journal of big Data*, 9(1):45.