From Dataset to Detection: A Comprehensive Approach to Combating Malayalam Fake News

Devika K[1], **Hari Prasath.S.B**[1], **Haripriya B**[1], **Vigneshwar E**[1], **Premjith B**[1], **Bharathi Raja Chakravarthi** [2]
[1]Amrita School of Artificial Intelligence, Coimbatore
Amrita Vishwa Vidyapeetham, India
[2]School of Computer Science, University of Galway, Ireland
b_premjith@cb.amrita.edu

Abstract

Identifying fake news hidden as real news is crucial to fight misinformation and ensure reliable information, especially in resource-scarce languages like Malayalam. To recognize the unique challenges of fake news in languages like Malayalam, we present a dataset curated specifically for classifying fake news in Malayalam. This fake news is categorized based on the degree of misinformation, marking the first of its kind in this language. Further, we propose baseline models employing multilingual BERT and diverse machine learning classifiers. Our findings indicate that logistic regression trained on LaBSE features demonstrates promising initial performance with an F1 score of 0.3393. However, addressing the significant data imbalance remains essential for further improvement in model accuracy.

1 Introduction

Detecting fake news is critical to identifying false or intentionally misleading information presented as legitimate news. In today's digital age, numerous websites spread fake news, significantly influencing society. The deceptive strategies employed by fake news to appear true further complicate this problem. Fake news has far-reaching consequences, shaping public opinion, interfering with democratic processes like elections, and even inciting violence. Researchers from various disciplines recognize the significance of studying and addressing this issue (Jain et al., 2019; Baarir and Djeffal, 2021; Choudhary et al., 2021).

Although technology and social media positively impact society, they are not without limitations or defects. The spread of fake news and the threat of inaccurate data have grown, potentially leading to serious social problems. The effects of fake news can be wide-ranging, from being merely annoying to influencing and misleading entire communities or even countries. Inaccurate information has a negative impact on society, according to related research. There are many ways to identify false news, including topic-agnostic, knowledge-based, machine-learning-based, and hybrid techniques.

The importance of classifying fake news in Malayalam and other low-resource languages lies in reducing the spread of false information and promoting informed decision-making in linguistically diverse societies. This requires developing effective models for classifying fake news in various languages, especially those with limited linguistic resources, such as Malayalam. Fake news frequently exploits linguistic and cultural particulars in low-resource languages, requiring the development of language-specific detection methods. Effective models for identifying false information in languages with limited resources, as demonstrated in (Raja et al., 2023b), (De et al., 2022), and (Nair et al., 2022), can play a crucial role in nurturing acceptance of diverse perspectives. Keeping fake news detection in low-resource languages relevant and effective in the digital age.

This paper makes a significant contribution to the field by providing a dataset and baseline machine learning models designed for classifying fake news into different classes based on the degree of misinformation it contains. To the best of our knowledge, this is the first dataset in this domain focusing on combating the spread of fake news to protect society from these inhumane acts of violence.

This paper delves into the strategies for detecting fake news and the complexities of distinguishing between the types of fake news. Deception, manipulation, and polarization are among the negative impacts that detection seeks to prevent by combating the spread of misleading information across various platforms, including social media and messaging apps. This pressing issue motivates us to dedicate our efforts to this work.

The absence of relevant data made it difficult to create a large corpus to identify and categorize

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 16–23

false news in Malayalam. Creating a sufficiently large and diverse dataset is more challenging for Malayalam, which is the topic of extensive research due to the absence of Malayalam-specific resources. The lack of data makes developing and evaluating reliable identification algorithms more challenging and highlights the urgent need for coordinated efforts to offer datasets related to Malayalam.

In addition, we explore various machine learning and deep learning algorithms to develop methods to detect fake information effectively. The morphological complexity and the complicated structures of Malayalam words and sentences made the feature extraction and classification model development a challenging task, needing the use of advanced techniques in order to capture the complex aspects of the language effectively. The state-of-the-art multilingual BERT models were utilized to transform the input sequence into embeddings, whereas Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Decision Tree (DT) and K-Nearest Neighbour (KNN) classification algorithms were employed for categorizing a fake content into different classes.

2 Literature review

(Raja et al., 2023b) employed two datasets for fine-tuning their pre-trained model. The first dataset is the English ISOT(Ahmed et al., 2018) dataset, consisting of actual and factual news articles. The second dataset is a new collection comprising regional languages such as Telugu, Kannada, Tamil, and Malayalam. For every news article represented by $(D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\})$, where x represents the news and y its corresponding label, the authors utilized a pre-trained mBERT or XLM-R model trained on a large, resourceful language dataset. This pre-trained model was fine-tuned for the new dataset D using an adaptive fine-tuning algorithm. The main approach involves transfer learning, taking a pre-trained model from one domain and adapting it to the target domain. The authors aimed to optimize the fine-tuning process and develop a model that maximizes the accuracy of detecting fake news in dataset D.

According to the study by (Raja et al., 2024), the Dravidian Languages, including Tamil, Telugu, Malayalam, and Kannada, have unique characteristics. The words are formed by adding affixes to the root word, making it challenging to find the meaning of the words without knowing the context in which they are used. The paper leverages the strengths of dilated temporal convolutional networks (DTCN) and integrates them with Bidirectional LSTM (BiLSTM) and a contextualized attention mechanism (CAM). The DTCN is employed to capture temporal dependencies, BiLSTM to seize long-range dependencies, and CAM to emphasize important information from the news while neglecting irrelevant content. The authors applied an adaptive cyclical learning rate with an early stopping mechanism to improve the model's convergence rate and accuracy. The dataset consists of news articles represented by (d_1, y_1), (d_2, y_2), ..., (d_n, y_n), where d represents the article and y its corresponding label. The researchers constructed a model to predict the label for each news article.

A Bala and P Krishnamurthy employed the MuRIL model in (Bala and Krishnamurthy, 2023) to detect fake news. MuRIL was refined by supervised learning on a handpicked dataset of labelled posts, comments, and keywords in Dravidian languages. The process of fine-tuning allowed the algorithm to distinguish between true and fake news. The MuRIL model examines textual information in each news to anticipate the classification and extracts semantic features. With the help of a sizable corpus of data from several Indian languages, MuRIL is a transformer-based architecture that has been pre-trained to capture linguistic subtleties and semantic correlations unique to the languages in the dataset.

(Balaji et al., 2023) proposed that data preprocessing is important for fake news detection since the appropriate form of data is required for training ML models. First data cleaning is done to eliminate punctuations, special characters and other HTML elements that don't add anything to the meaning of the news. The text is then separated into distinct words to produce a structured representation. Words are shortened to their base form using stemming processes to maintain consistency. Finally the text is vectorized before feeding it to the machine learning model. Various models like the BERT, ALBERT, XLNET, M-BERT are used in this paper and M-BERT comes out on top with an accuracy of 0.74. M-BERT is a version of BERT which supports multilingual text feasibly. It is trained on a combination of monolingual and multilingual data by which it gains the ability to produce language representations of languages from different origins. Fine-tuning the model on task-specific labelled data across many languages is a necessary step.

(Coelho et al., 2023) deployed fake news detection models for Malayalam. In this work, the punctuations and special characters were removed in data pre-processing and the text is converted into its equivalent English form which is useful for classification. An ensemble machine learning classifier was proposed in this paper to identify fake news.

(Raja et al., 2023a) developed a XLM-R model for fake detection in Malayalam. The XLM-RoBERTa model is also a multilingual variant of the BERT based transformer model. It consist of self-attention mechanisms which enables it to learn contextualized word embeddings which helps in capturing relationships between words in a sentence, ultimately tuning the model to encode the semantic information of the input text effectively. The model is fine-tuned using an annotated Malayalam fake news dataset. It allows the model to learn specific patterns and linguistic characteristics of fake news in Malayalam. The news is labelled genuine or fake by augmenting the model with a classification layer on the top. The parameters of the model are updated during the fine-tuning process. Bayesian optimizer wass used to find the optimal hyperparameters for the deep learning based model which maximizes the model's performance. The proposed XML-RoBERTa model achieved a F1-Score of 87%.

According to (Oshikawa et al., 2020), the fake news detection problem is often viewed as a classification problem rather than a regression problem since regression gives us an output of a numeric score of the integrity of the data. Pre-processing steps followed in this work includes tokenization, stemming and weighting of words. The input texts were converted into features using TF-IDF. Though various Machine Learning models were used for fake news detection, the Neural Network based model achieved the highest accuracy in detection. Attention mechanisms are incorporated into neural networks to boost their performance and accuracy.

(Thota et al., 2018) implemented three different variations of neural networks. The first model utilizes TF-IDF with Dense Neural Networks. This model takes the TF-IDF vector of the headline pair's cosine similarity, a standard practice to measure similarity between non-zero vectors, as input and predicts the output. The vectors are passed to the dense network layers, and the final dense layer predicts the output label for the text news. The second model employs a Bag of Words vector with Deep Neural Networks (DNN). It uses a simplified vector space embeddings to represent text. The third model incorporated a pre-trained word embedding model trained using neural networks. For this neural network architecture, Word2Vec was employed to represent words in a 300-dimensional vector space, and these embeddings are fed into the classification model. Among these, the TF-IDF-based model was the best-performing model. To address the potential overfitting of the neural network model, various regularization techniques such as L2 and early stopping, are deployed to improve generalization. This model has demonstrated superior performance compared to existing model architectures, achieving an accuracy of 94% on the test data.

3 Fake news dataset in Malayalam

Even though there are datasets available for checking whether news is fake, there are no datasets available in Malayalam to check how much misinformation a news carries, which motivated our research. Therefore, we refer to various fact-checking websites to prepare an authentic dataset to measure different levels of misinformation in the news. This process posed different challenges.

- **The selection of the fact-checking websites**: We addressed this issue by selecting the fact-checking pages of the mainstream media in Malayalam. The list of websites from where the data was collected are listed below
  - Newsmeter [1]
  - India Today Malayalam [2]
  - Malayalam Factcrescendo [3]
  - Asianet News [4]
- **Annotation**: Instead of manually annotating each piece of news, we select the labels provided by the fact-checking websites. This work aims to classify the different classes of fake news. Instead of labelling news as either true or fake, we decided to collect the news, which is categorized into different degrees of falseness. The labels we used to classify are False, Partly False, Mostly False, and Half True. We labelled the news as False when the entire news is untrue, Partly False when

[1] https://newsmeter.in/fact-check-malayalam
[2] https://malayalam.indiatoday.in/fact-check/
[3] https://www.malayalam.factcrescendo.com/
[4] https://www.asianetnews.com/fact-check

Classes	Train set	Test set
Half True	145	24
False	1,251	149
Partly False	44	14
Mostly False	242	63
Total	1,682	250

Table 1: Class-wise distribution of the dataset

the news contains a mixture of accurate and inaccurate information with a small portion being false, Mostly False when the news is false but contains some true information, and Half True when the news has equal true and untrue information.

- **The authenticity of the annotation**: To ensure that the labels are not biased, we cross-checked the fake category of each piece of news with different fact-checking websites.
- **Identification of the fake news**: Instead of going through every fact-checking website, we searched specific keywords and collected the fact-checking results.
- **Redundancy in data**: Data was repeated as the news was collected from various sources. The sources were limited; some were labelled as misleading rather than clearly true, partially true, or false.
- **Morphological complexity of Malayalam words**: The morphological richness and complexity of the Malayalam language made it challenging to identify keywords for news retrieval. We had to search for different word forms to collect different news sets about one particular keyword.

The collected dataset was divided into train data and test data. The class-wise statistics of the train and test datasets are given in Table 1.

4 Methodology

The Block diagram of the proposed model is shown in Figure 1 and the steps involved in the proposed methodology are described below.

The classification of fake news into distinct subcategories was modelled as a text classification problem in which a sequence of tokens served as the input, and a class label representing a fake news category was considered the output. We employed BERT-based multilingual models to generate embeddings for the input token sequence. Besides, machine learning classifiers were utilised to determine the function that converts the embedding into the appropriate labels.

The collected data was divided into training and test sets. Both training and test datasets were converted into vector representation using BERT-based models. We used multilingual sentence transformers for transforming input text sequences into embeddings. The resultant features and their labels were used for building the classifier. Since the number of data points is less in the training data, we decided to implement a cross-validation-based grid search approach to fix the optimal hyperparameters of each classification model. The best estimators were used to train the model using the training data.

5 Experiments

This section provides an overview of the five experiments conducted. We considered four multilingual BERT models and one Malayalam-specific BERT model for generating the embeddings. The input sequences were fed into the sentence transformer designed using the abovementioned BERT models to generate the feature vector. The multilingual BERT models considered for this work are - BERT multilingual-cased, MuRIL, LaBSE, and indicBERT. The classification of news into different categories was modelled using five different classification algorithms - Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Logistic Regression (LR) and K-Nearest Neighbor (KNN) (Premjith et al., 2019). The following subsections explain the performance of different classifiers with each embedding model.

The first experiment uses the Malayalam BERT model (Joshi, 2022). This model is trained on publicly available Malayalam monolingual datasets. The performance of each classifier with the embeddings generated using the Malayalam BERT model is given in Table 2.

The training dataset is highly imbalanced, and most of the data belong to the FALSE category. This imbalance may force the model to be biased towards the majority class. Therefore, accuracy cannot be trusted as the best metric to evaluate the performance of a classification model. Consequently, we considered the macro F1 score as the metric to assess the prediction capabilities of each classifier. Table 2 shows that RF exhibited the high-

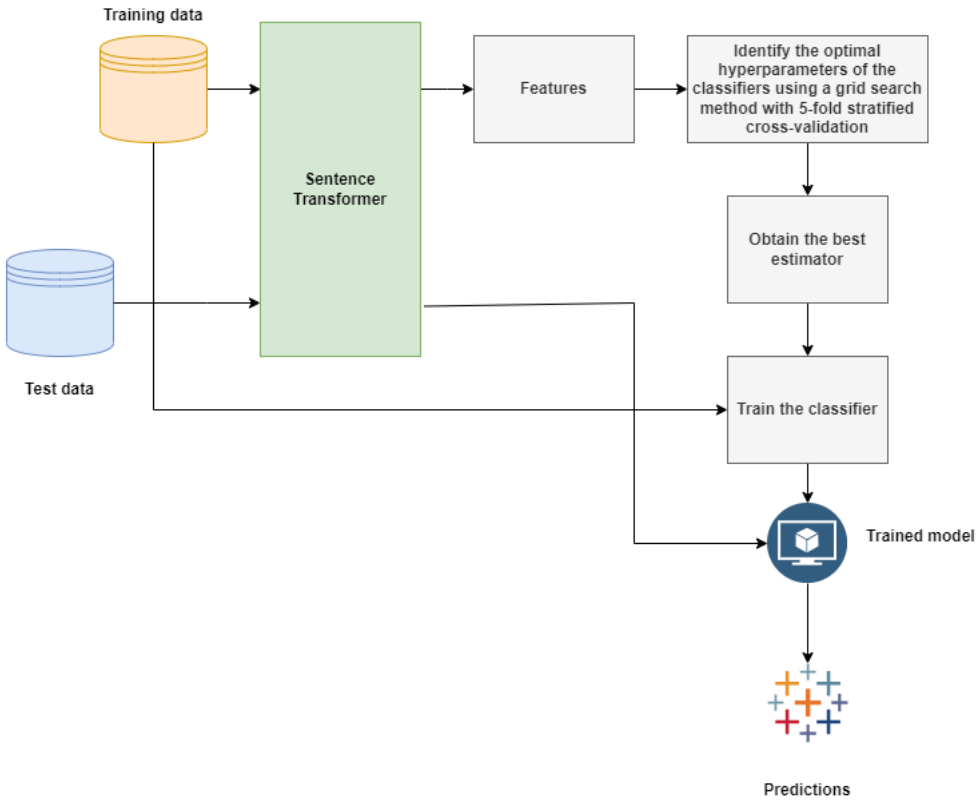

Figure 1: Flow diagram for the proposed fake news classification methodology

Classifier	Accuracy	Precision	Recall	F1 score
SVM	63.69	0.4154	0.3253	**0.3392**
RF	**74.40**	0.4358	0.2551	0.2231
LR	66.39	0.3046	0.3070	0.3052
DT	64.88	0.3235	0.3018	0.3094
KNN	72.32	0.3117	0.2638	0.2497

Table 2: Performance of the malayalam-bert model

Classifier	Accuracy	Precision	Recall	F1 score
SVM	**74.10**	0.1852	0.2500	0.2128
RF	**74.10**	0.1852	0.2500	0.2128
LR	65.17	0.3374	0.3346	**0.3353**
DT	63.69	0.2653	0.2675	0.2658
KNN	72.61	0.2615	0.2572	0.2360

Table 3: Performance of the multilingual-cased model

Classifier	Accuracy	Precision	Recall	F1 score
SVM	63.98	0.2775	0.2843	0.2799
RF	**74.40**	0.4358	0.2551	0.2231
LR	68.75	0.3310	0.3162	**0.3156**
DT	66.66	0.2875	0.2869	0.2846
KNN	73.51	0.2712	0.2561	0.2312

Table 4: Performance of the MuRIL model

Classifier	Accuracy	Precision	Recall	F1 score
SVM	**74.10**	0.1852	0.2500	0.2128
RF	**74.40**	0.4358	0.2586	0.2298
LR	67.55	0.3542	0.3308	**0.3393**
DT	62.79	0.2910	0.2909	0.2907
KNN	72.61	0.2784	0.2613	0.2442

Table 5: Performance of the LaBSE model

Classifier	Accuracy	Precision	Recall	F1 score
SVM	63.98	0.2789	0.2796	**0.2786**
RF	**74.40**	0.4358	0.2551	0.2231
LR	66.36	0.2850	0.2795	**0.2786**
DT	64.88	0.2644	0.2627	0.2600
KNN	71.72	0.2321	0.2501	0.2266

Table 6: Performance of the IndicBERT model

est accuracy of 74.40%, followed by KNN with an accuracy of 72.32%. However, SVM achieved the maximum F1 score of 0.3392. In the case of RF and KNN, the precision was higher than recall, which means that the model was more accurate in predicting the positive class but failed to capture all other relevant results. The optimal hyperparameters for building the SVM model are $C = 0.1, gamma = 1, kernel = linear$. The RF classifier was built using 200 estimators.

In the second experiment, we used the multilingual BERT base-cased model (Devlin et al., 2018) to compute the vector representation for the input data. This model is pre-trained on a large corpus of multilingual data in a self-supervised fashion. It is pre-trained with data collected from 104 languages. The performance of each classifier is shown in Table 3.

The best-performing classifiers in terms of accuracy are SVM and RF. Both models achieved an accuracy of 74.10%. Nevertheless, LR demonstrated the best performance with multilingual BERT-based features regarding the F1 score with an F1 score of 0.3353. The optimal hyperparameters for developing the LR model were $C = 0.1, penalty = L2$, whereas SVM and RF were built using the parameters $C = 10, gamma = 0.1, kernel = RBF$ and $n_estimators = 50$, respectively.

The third experiment used the MuRIL (Khanuja et al., 2021) model for generating the embedding. This model is pre-trained using 17 Indian languages. This model is pre-trained on translation and their transliterated counterparts. Table 4 describes the performance of different classifiers in categorizing the news into different classes with MuRIL embeddings. In this experiment, RF attained the highest accuracy of 74.40%, and LR exhibited the best F1 score of 0.3156. The RF model was trained using 100 estimators, whereas we considered the hyperparameters $C = 0.1, penalty = L2$ for training the model.

In the fourth experiment, the LaBSE model was used to transform the input text into embeddings. This model is trained and optimized for bilingual sentence pairs. The performance scores of different classifiers used in this experiment are shown in Table 5. Here, both SVM and RF showed the best accuracy with a score of 74.10%, whereas LR achieved the best F1 score of 0.3353. The SVM, RF and LR were trained using the hyperparameters $C = 1, gamma = 1, kernel = RBF$, 50 estimators and $C = 0.1, penalty = L2$, respectively.

In the fifth experiment, we utilized IndicBERT (Kakwani et al., 2020), a multilingual model pretrained on 12 major Indian languages. The result of this experiment is shown in Table 6. It is observed that RF obtained an accuracy of 74.40%, and SVM and LR scored the highest F1 score of 0.2786. RF model consisted of 50 estimators, whereas SVM and LR models were built using the hyperparameters $C = 0.1, gamma = 1, kernel = Linear$ and $C = 0.1, penalty = L2$, respectively.

Among all the classifiers, LR achieved the highest F1 score with four feature embeddings, and SVM demonstrated the best performance with Malayalam BERT, with a score of 0.3392. The

best score attained by LR was 0.3393 LaBSE embeddings. LR exhibited comparable performance with an F1 score of 0.3052 with Malayalam BERT embeddings.

6 Conclusion

This paper proposes a new dataset to categorize fake news in Malayalam into different fake categories based on the degree of falseness. To the best of our knowledge, this is the first dataset curated for fake news classification in Malayalam. In addition, we developed baseline models for identifying the fake category of false news in this work using various multilingual BERT models and machine learning classifiers. Among all the models, the logistic regression model trained over LaBSE features was the best, with an F1 score of 0.3393. The high imbalance in the training data significantly affected the model's performance.

The dataset exhibits a significant imbalance, potentially resulting in model biases favouring the majority class. Potential solutions to this problem include implementing cost-sensitive learning and oversampling techniques; these represent the future trajectories of this research.

Acknowledgements

The author Bharathi Raja Chakravarthi was supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2(Insight_2).

References

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2018. Detecting opinion spams and fake news using text classification. *SECURITY AND PRIVACY*, 1(1):e9.

Nihel Fatima Baarir and Abdelhamid Djeffal. 2021. Fake news detection using machine learning. In *2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH)*, pages 125–130.

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.

Murari Choudhary, Shashank Jha, Deepika Saxena, and Ashutosh Singh. 2021. A review of fake news detection methods using machine learning.

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.

Arkadipta De, Dibyanayan Bandyopadhyay, Baban Gain, and Asif Ekbal. 2022. A transformer-based approach to multilingual fake news detection in low-resource languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 21:1–20.

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

Anjali Jain, Avinash Shakya, Harsh Khatter, and Amit Gupta. 2019. A smart system for fake news detection using machine learning. pages 1–4.

Raviraj Joshi. 2022. L3Cube-HindBERT and DevBERT: Pre-Trained BERT Transformer models for Devanagari based Hindi and Marathi languages. *arXiv preprint arXiv:2211.11418*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

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.

Jayashree Nair, S S Akhil, and V Harisankar. 2022. Fake news detection model for regional language. In *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, pages 1–7.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A survey on natural language processing for fake news detection.

B Premjith, KP Soman, M Anand Kumar, and D Jyothi Ratnam. 2019. Embedding linguistic features in word embedding for preposition sense disambiguation in English—Malayalam machine translation context. *Recent advances in computational intelligence*, pages 341–370.

Eduri Raja, Badal Soni, and Sami Kumar Borgohain. 2023a. nlpt malayalm@DravidianLangTech : Fake news detection in Malayalam using optimized XLM-RoBERTa model. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 186–191, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Eduri Raja, Badal Soni, and Samir Kumar Borgohain. 2023b. Fake news detection in Dravidian languages using transfer learning with adaptive finetuning. *Engineering Applications of Artificial Intelligence*, 126:106877.

Eduri Raja, Badal Soni, Candy Lalrempuii, and Samir Kumar Borgohain. 2024. An adaptive cyclical learning rate based hybrid model for Dravidian fake news detection. *Expert Systems with Applications*, 241:122768.

Aswini Thota, Priyanka Tilak, Simrat Ahluwalia, and Nibrat Lohia. 2018. Fake News Detection: A Deep Learning Approach. *SMU Data Science Review*, 1(3):10.