# AbhiPaw@ DravidianLangTech: Fake News Detection in Dravidian Languages using Multilingual BERT

**Abhinaba Bala**
IIIT Hyderabad, India
abhinaba.bala@research.iiit.ac.in

**Parameswari Krishnamurthy**
IIIT Hyderabad, India
param.krishna@iiit.ac.in

## Abstract

This study addresses the challenge of detecting fake news in Dravidian languages by leveraging Google's MuRIL (Multilingual Representations for Indian Languages) model. Drawing upon previous research, we investigate the intricacies involved in identifying fake news and explore the potential of transformer-based models for linguistic analysis and contextual understanding. Through supervised learning, we fine-tune the "muril-base-cased" variant of MuRIL using a carefully curated dataset of labeled comments and posts in Dravidian languages, enabling the model to discern between original and fake news. During the inference phase, the fine-tuned MuRIL model analyzes new textual content, extracting contextual and semantic features to predict the content's classification. We evaluate the model's performance using standard metrics, highlighting the effectiveness of MuRIL in detecting fake news in Dravidian languages and contributing to the establishment of a safer digital ecosystem.

Keywords: fake news detection, Dravidian languages, MuRIL, transformer-based models, linguistic analysis, contextual understanding.x

## 1 Introduction

The proliferation of fake news has become a pressing concern in the digital age, significantly impacting the accuracy and credibility of information dissemination. Detecting and combating fake news is a critical task to ensure the integrity of news sources and promote informed decision-making. While extensive research has been conducted in various languages and domains, the detection of fake news in the Dravidian language poses unique challenges due to its linguistic characteristics and cultural nuances.

The challenges involved in fake news detection have been widely explored in the literature. Previous studies have examined different task formulations, datasets, and NLP solutions to address this issue (Oshikawa et al., 2020). Understanding the potential and limitations of existing approaches is essential for developing effective strategies in the context of the Dravidian language.

While existing research has made significant contributions to fake news detection, the unique linguistic characteristics and cultural context of the Dravidian language necessitate dedicated efforts and language-specific approaches. This paper aims to address the challenges of fake news detection in the Dravidian language by exploring methodologies, datasets, and potential avenues for future research. By leveraging insights from related work, we aim to contribute to the development of effective and culturally sensitive solutions for detecting fake news in the Dravidian language space.

## 2 Related Work

To understand the challenges involved in fake news detection (Oshikawa et al., 2020) made a survey, along with describing the related tasks. They systematically review and compare the task formulations, datasets and NLP solutions that have been developed for the task, and also discuss the potentials and limitations of them. (Goldani et al., 2020) propose a new model based on capsule neural networks for detecting fake news, they also propose two architectures for different lengths of news statements.

With increase in popularity of graph networks, following are some works in fake news detection. (Veselý and Veselý, 2021) propose FANG, a novel graphical social context representation and learning framework for fake news detection. Unlike previous contextual models that have targeted performance, their focus is on representation learning. (Ren et al., 2020) present Adversarial Active Learning-based Heterogeneous Graph Neural Net-

*Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 235–238, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_034

work (AA-HGNN) which employs a novel hierarchical attention mechanism to perform node representation learning in the heterogeneous information network HIN.

(Liu et al., 2020) show that given a claim and a set of potential evidence sentences that form an evidence graph, Kernel Graph Attention Network (KGAT) introduces node kernels, which better measure the importance of the evidence node, and edge kernels, which conduct fine-grained evidence propagation in the graph, into Graph Attention Networks for more accurate fact verification. For more realisitc scenarios and social media (Lu and Li, 2020) aim at predicting whether a source tweet is fake or not, and generating explanation by highlighting the evidences on suspicious retweeters and the words they concern. They develop a neural network-based model, Graph-aware Co-Attention Networks (GCAN), to achieve the goal.

(Sabarmathi et al., 2021) make two contributions - a. present two new datasets for the undertaking of fake information identification which covers several domains, b. test and train a set of mastering discoveries to create precise fake news detectors.

(Lucas et al., 2022) attempt to detect COVID-19 misinformation (in English, Spanish, and Haitian French) populated in the Caribbean regions. They trained several classification and language models on COVID-19 in the high-resource language regions and transferred the knowledge to the Caribbean claim dataset.

Ensembling machine learning models are quite competant as depicted by - (Akram and Shahzad, 2021), (Kalraa et al., 2021). (Akram and Shahzad, 2021) employ a voting-based approach of the three most effective techniques to decide that the given news article is fake or real. They perform experiments using several classical machine learning techniques, three types of features, unigram, bigram and trigram.

(Kalraa et al., 2021) use ensemble of Various Transformer Based Models for the Fake News Detection Task in the Urdu Language. (Ameer et al., 2021) use transfer learning with BERT algorithm. (Lin et al., 2020) use CharCNN along with Roberta to obtain sentence embeddings with respect to both the word level and character level. They adopt label smoothing to improve the generalization capability of the model.

(Sivanaiah et al., 2023) prepare fake news dataset for low resource languages (Tamil, Kannada, Gujarati, and Malayalam); they experiment with Logistic Regression and BERT models to perform the detection. (De et al., 2021) also offer a multilingual multidomain fake news detection dataset of five languages and seven different domains for resource scarce scenarios. They propose an effective neural model based on the multilingual Bidirectional Encoder Representations from Transformer (BERT) for domain-agnostic multilingual fake news classification.

## 3 Method

The task is to classifying YouTube comments as either original or fake news.

### 3.1 Classifier and Feature Extraction

We employed Google's MuRIL (Multilingual Representations for Indian Languages) model, specifically the ”muril-base-cased” variant, for our classification task. MuRIL is a state-of-the-art language representation model designed for Indian languages, based on the transformer architecture. It has been pre-trained on a large corpus of text data from various Indian languages, capturing linguistic nuances and semantic relationships specific to these languages.

To adapt MuRIL to our classification task, we fine-tuned the ”muril-base-cased” model on our dataset using supervised learning. Labeled examples of YouTube comments were used off the dataset, with labels indicating whether they were original or fake news. This fine-tuning process allowed MuRIL to learn relevant features and patterns for distinguishing between original and fake news in our specific context.

Using the fine-tuned MuRIL model, we performed classification on new, unseen comments or posts. MuRIL computed representations that captured contextual information and semantic meaning, and a classification layer was applied to predict the category of each comment or post (original or fake news. F1 score was used as the evaluation metric to assess MuRIL's performance on our classification task.

### 3.2 Training and Inference

Through supervised learning, MuRIL adjusts its parameters to minimize the classification loss and optimize its performance on the training data. By fine-tuning on the task-specific dataset, MuRIL learns to capture relevant features and patterns for

Table 1: Data Distribution

| Category | Count |
|---|---|
| Fake | 1599 |
| Original | 1658 |

distinguishing between original and fake news.

In the inference phase, the fine-tuned MuRIL model is applied to unseen or new examples to make predictions. Given a comment or post, MuRIL computes representations that capture contextual information and semantic meaning. These representations are then fed into a classification layer, which predicts the category of the input (original or fake news). The predictions are based on the learned patterns and features obtained during the training phase. In this way, MuRIL performs classification on new instances, providing accurate and reliable predictions based on its understanding of the data it has been trained on.

## 4 Experiments and Results

### 4.1 Evaluation Metrics

The macro average F1-score is the performance metric used to evaluate the overall effectiveness of the detection model. It is derived by calculating the F1-score for each individual class and then taking the average across all classes. Regardless of class size or class imbalance, the macro average F1-score considers the performance of each class independently and then computes the average, giving equal importance to all classes.

### 4.2 Datasets

We perform our experiments on ”Youtube comments in the Malayalam language” annotated for fake news detection. The dataset is shared as a part of DravidianLangTech@RANLP 2023.

#### 4.2.1 Dataset Analysis

**Class Imbalance**

As evident from the distribution of labels, there is no class imbalance (Table 1).

### 4.3 Results

We achieved a macro F1-score of 0.87 which placed us second in the leader-board, Table 2.

Table 2: AbhiPaw @ Fake News Detection in Dravidian Languages

| Team | F1-score (macro) | Rank |
|---|---|---|
| DeepBlueAI-alert | 0.9 | 1 |
| AbhiPaw | 0.87 | 2 |
| nit-it-nlp | 0.87 | 2 |
| nlpt | 0.87 | 2 |
| MUCS | 0.83 | 3 |
| ML-AL-IIIT-Ranch | 0.78 | 4 |
| DLRG-RR | 0.73 | 5 |
| NLP-SSN-CSE | 0.73 | 6 |

## 5 Conclusion

In conclusion, we address the crucial issue of fake news detection in Dravidian languages by harnessing the power of Google's MuRIL (Multilingual Representations for Indian Languages) model. Through fine-tuning and supervised learning, we successfully trained MuRIL to accurately classify comments and posts as original or fake news. The experimental results demonstrate the efficacy of MuRIL in capturing linguistic nuances and contextual information specific to Dravidian languages, enhancing the model's ability to detect fake news. Future research directions include expanding the dataset to improve generalization, addressing code-switching and domain adaptation challenges, exploring ensemble methods, detecting subtle forms of fake news, and extending the research to other low-resource languages for cross-lingual transfer learning.

## References

Hammad Akram and Khurram Shahzad. 2021. Ensembling machine learning models for urdu fake news detection. In *Fire*.

Iqra Ameer, Claudia Porto Capetillo, Helena Gómez-Adorno, and Grigori Sidorov. 2021. Automatic fake news detection in urdu language using transformers. In *Fire*.

Arkadipta De, Dibyanayan Bandyopadhyay, Baban Gain, and Asif Ekbal. 2021. A transformer-based approach to multilingual fake news detection in low-resource languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(1).

Mohammad Hadi Goldani, Saeedeh Momtazi, and Reza Safabakhsh. 2020. Detecting fake news with capsule neural networks. *Appl. Soft Comput.*, 101:106991.

Sakshi Kalraa, Preetika Vermaa, Yashvardhan Sharma, and Gajendra Singh Chauhan. 2021. Ensembling of various transformer based models for the fake news detection task in the urdu language. In *Working Notes* *of FIRE 2021 - Forum for Information Retrieval Evaluation, Gandhinagar, India, December 13-17, 2021*. CEUR-WS.org.

Nankai Lin, Sihui Fu, and Shengyi Jiang. 2020. Fake news detection in the urdu language using charcnn-roberta. In *Fire*.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351, Online. Association for Computational Linguistics.

Yi-Ju Lu and Cheng-Te Li. 2020. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 505–514, Online. Association for Computational Linguistics.

Jason Lucas, Limeng Cui, Thai Le, and Dongwon Lee. 2022. Detecting false claims in low-resource regions: A case study of Caribbean islands. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 95–102, Dublin, Ireland. Association for Computational Linguistics.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A survey on natural language processing for fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6086–6093, Marseille, France. European Language Resources Association.

Yuxiang Ren, Bo Wang, Jiawei Zhang, and Yi Chang. 2020. Adversarial active learning based heterogeneous graph neural network for fake news detection. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 452–461.

K R Sabarmathi, K Gowthami, and S Sanjay Kumar. 2021. Fake news detection using machine learning and natural language inference (nli). *IOP Conference Series: Materials Science and Engineering*, 1084(1):012018.

Rajalakshmi Sivanaiah, Nishaanth Ramanathan, Shajith Hameed, Rahul Rajagopalan, Angel Deborah Suseelan, and Mirnalinee Thanka Nadar Thanagathai. 2023. Fake news detection in low-resource languages. In *Speech and Language Technologies for Low-Resource Languages*, pages 324–331, Cham. Springer International Publishing.

Dominik Veselý and Marek Veselý. 2021. Reproducibility study - fang: Leveraging social context for fake news detection using graph representation.