BlueRay@DravidianLangTech-2025: Fake News Detection in Dravidian Languages

Kogilavani Shanmugavadivel[1], **Malliga Subramanian**[2],
Aiswarya M[1], **Aruna T**[1], **Jeevaananth S**[1]
[1]Department of AI, Kongu Engineering College, Perundurai, Erode.
[2]Department of CSE, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{aiswaryam.22aid, arunat.22aid}@kongu.edu
{jeevaananths.22aid}@kongu.edu

Abstract

The rise of fake news presents significant issues, particularly for underrepresented languages. This study tackles fake news identification in Dravidian languages with two subtasks: binary classification of YouTube comments and multi-class classification of Malayalam news into five groups. Text preprocessing, vectorization, and transformer-based embeddings are all part of the methodology, including baseline comparisons utilizing classic machine learning, deep learning, and transfer learning models. In Task 1, our solution placed 17th, displaying acceptable binary classification performance. In Task 2, we finished eighth place by effectively identifying nuanced categories of Malayalam news, demonstrating the efficacy of transformer-based models.

1 Introduction

Fake news detection is a critical difficulty in combatting disinformation in today's digital landscape. Fake news is defined as intentionally misleading or incorrect material presented as legitimate news, which is typically designed to confuse readers and alter public opinion said by Anitha et al. (2024). In the view of Subramanian et al. (2025) proliferation of digital media has increased the dissemination of fake news, allowing misinformation to reach a large audience. Bala and Krishnamurthy (2023) highlight that fake news can take many forms, including manufactured tales, altered media, and biased content, especially on social media platforms where false narratives can quickly spread.

Devika et al. (2024) argue that misinformation adds to public panic, political polarization, and a reduction in faith in trustworthy news sources. Furthermore, unregulated fake news can sway public opinion, affect elections and policymaking, and incite social upheaval. Hariharan and Anand Kumar (2022) says detecting fake news is difficult due to the variety of writing styles, linguistic difficulties, and false news' ability to replicate actual information. Mohan et al. (2024) emphasize that standard detection methods frequently fail to capture contextual and cultural nuances, necessitating advanced natural language processing (NLP) models customized to these languages. According to Bade et al. (2024), machine learning and deep learning approaches are vital for developing robust false news detection models.

The shared task Fake News Detection in Dravidian Languages [1]aims on detecting fake news in underrepresented languages using binary and multi-class classification sets. This study describes a system for detecting fake news in various settings that uses text preprocessing, vectorization techniques (TF-IDF, BERT, etc.), advanced classification models such as transformers, and classic machine learning approaches. Section 2 summarizes works on detecting fake news, whereas Section 3 provides a full system description. Section 4 presents experimental results and analysis, followed by insights and future research directions.

2 Literature Review

Several studies have explored fake news detection in Dravidian languages, particularly Malayalam and Tamil, using various machine learning and deep learning approaches like Subramanian et al. (2023). Raja et al. (2023) proposed an optimized XLM-RoBERTa model, achieving improved accuracy in Malayalam fake news detection. Similarly, Sujan et al. (2023) introduced MalFake, a multimodal framework integrating Recurrent Neural Networks (RNNs) and VGG-16, demonstrating the effectiveness of combining text and images for misinformation identification. Coelho et al. (2023) adopted a traditional machine learning approach, experimenting with different classifiers for fake news detection. Eduri et al. (2023) ex-

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

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 313–318

plored gradient accumulation-based transformer models, improving fake news classification performance in Malayalam. Additionally, Subramanian et al. (2024) provided an overview of the second shared task on fake news detection, highlighting key methodologies and benchmark datasets for Dravidian languages.

Other research efforts have focused on related NLP tasks for Malayalam and Tamil. Rameesa and Veeramanju conducted a systematic review on news headline categorization in Malayalam, addressing challenges in linguistic variations. Kumar et al. (2019) implemented deep learning-based part-of-speech tagging for Malayalam Twitter data, showcasing the importance of morphological analysis in NLP tasks. Ponnusamy et al. (2024) introduced an annotated dataset for misogyny detection in Tamil and Malayalam memes, emphasizing the role of social media in the spread of harmful narratives. Furthermore, YP and Nelliyullathil (2023) studied the spread of misinformation on Facebook, analyzing user engagement and the effectiveness of third-party fact-checking in curbing fake news. Farsi et al. (2024) improved MuRIL BERT, a multilingual BERT model designed for Indian languages, to classify fake news in Malayalam, with encouraging results. Rahman et al. (2024) used Malayalam-BERT, a language-specific transformer model, to classify fake news. They emphasized the relevance of domain-specific embeddings in increasing classification accuracy.

These studies collectively highlight the growing interest in fake news detection and NLP tasks in Dravidian languages Madhumitha et al. (2024). The advancements in transformers, multimodal learning, and traditional ML techniques have significantly contributed to improving detection accuracy, while challenges in code-mixing, linguistic diversity, and limited annotated datasets remain key areas for future research Osama et al. (2024).

3 Problem and System Description

The propagation of fake news on digital platforms has become a serious concern, fueling misinformation and upseting societal cohesion. This problem becomes more acute in multilingual populations, where code-mixed content hamper identification methods. Addressing this issue is critical to maintaining the credibility of the information shared online.

3.1 Dataset Description

The shared task dataset includes two subtasks with distinct structures:

Subtask 1 (Binary Classification): For this task the dataset has columns text and label. Column text refers to YouTube comments posted in Malayalam-English and label indicates if the comment is original or fake.

Subtask 2 (Multiclass Classification): For this task the dataset has columns Id, News, and Label. The Id is a unique identification given to each news story. Column News includes Malayalam news articles. Label sorts the news into five categories. The dataset is partitioned into two sets: training and testing.

Subtasks	Train	Test
Task 1	3,258	1020
Task 2	1901	200

Table 1: Dataset Description

3.2 Development Pipeline

Our system uses a systematic pipeline to detect fake news, which consists of the following stages: Text preprocessing, feature extraction, classification models, evaluation metrics. Figure 1 shows the workflow to detect fake news.

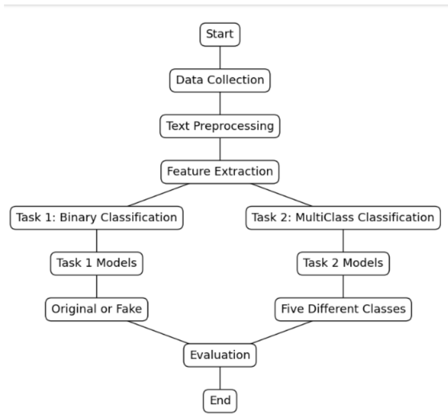


Figure 1: Proposed System Workflow.

3.2.1 Text Preprocessing

Text preparation is essential while creating Malayalam-English code-mixed YouTube comments and Malayalam news articles to detect fake news. To clean and organize the data efficiently, several techniques were required. When working with mixed-script tokens, the text was broken into words. Lowercasing and script normalisation ensured homogeneity. Stopwords and noise were removed using regular expression patterns, which included words, mentions, hashtags, emojis, and special characters. To preserve semantic meaning, words were stemmed and lemmatized in Malayalam to their base forms using language-specific procedures. Vectorization entailed transforming text into numerical representations using TF-IDF, Word2Vec, and transformer-based embeddings (BERT).

These preprocessing strategies ensured that models focused on meaningful content while decreasing noise and redundancy, resulting in higher classification accuracy.

3.2.2 Feature Extraction

Feature extraction translates text data into meaningful numerical representations, allowing for more successful fake news categorization. TF-IDF (Term Frequency-Inverse Document Frequency) emphasizes essential words while decreasing the influence of frequently used terms. Word Embeddings (Word2Vec) capture semantic links between words to improve contextual understanding, particularly in code-mixed text. Transformer-Based Embeddings (BERT) offers deep contextual meaning, improving classification accuracy for multilingual content.

These strategies aid the model's ability to discover patterns in both fake and authentic news, hence enhancing performance.

3.2.3 Classification Models

To efficiently recognize fake news in Dravidian languages, we used a variety of machine learning, deep learning models and transfer learning methods with various feature extraction methods designed to address the unique challenges of each task. Each model is briefly explained here, along with its performance.

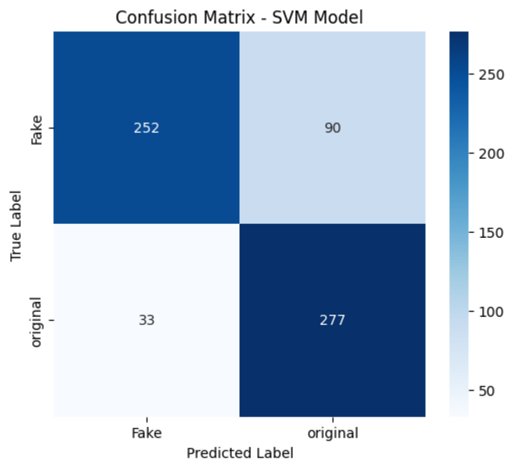
Task 1: Binary Classification (Fake vs Original in Code-Mixed Youtube Comments)

SVM with TF-IDF: This model uses an optimal decision boundary to distinguish between fake and original news. Gradient Boosting Classifier with TF-IDF: This sequential learning approach corrects prior errors while detecting complicated patterns in false news. Logistic Regression with CountVectorizer: It trains the model using word frequency representation, resulting in successful text classification based on term occurrence patterns. Random Forest Classifier with Word2Vec: Uses word embeddings to capture semantic meaning, which improves classification accuracy.

Classification Model	Accuracy
SVM with TF-IDF	0.81
Gradient Boosting Classifier with TF-IDF	0.80
Logistic Regression with CountVectorizer	0.77
Random Forest Classifier with Word2Vec	0.65

Table 2: Accuracy of Binary Classification Models (Task 1).

The accuracy gained by these models is displayed in table 2 and the figure 2 shows the performance of SVM with TF-IDF model.



1

Figure 2: Performance of SVM with TF-IDF Model.

Task 2: Multiclass Classification (Classifying Malayalam News into Fake News Types)

Bi-LSTM: A deep learning model that extracts contextual meaning from both past and future words, improving classification accuracy for false news categories. XGBoost Classifier: It is a powerful boosting method that can handle imbalanced datasets and learn complex word associations. DistilBERT: Improves text comprehension through transformer-based contextual embeddings, resulting in high accuracy in fake news classification. SVM for Multiclass: Extends SVM for multiclass classification by specifying the boundaries between news categories.

The accuracy gained by these models is displayed in Table 3 and the figure 3 shows the performance of DistilBERT model.

Classification Model	Accuracy
DistilBERT	0.68
SVM for Multiclass	0.67
XGBoost Classifier	0.64
Bi-LSTM	0.54

Table 3: Accuracy of Multiclass Classification Models (Task 2).

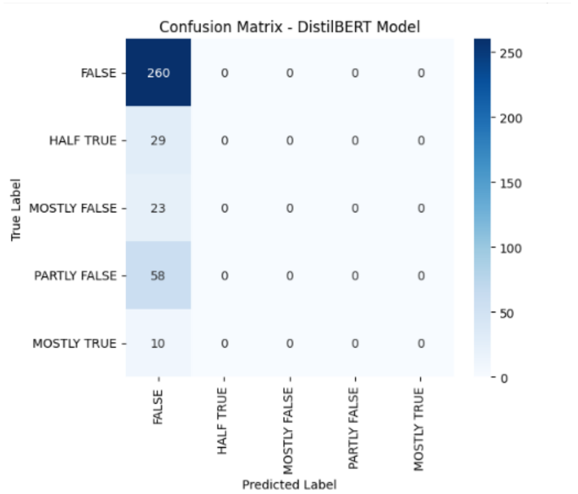


Figure 3: Performance of DistilBERT Model.

3.2.4 Evaluation Metrics

To ensure reliable fake news detection, the models are tested for accuracy, precision, recall, F1-score, macro F1-score, and loss. These measures helps determining the model's efficiency.

4 Experiments and Results

To classify fake news, experiments are conducted using various machine learning and deep learning models. For Task 1 (binary classification), SVM with TF-IDF attained a highest accuracy of 0.81 using a linear kernel, C value of 1.0, and balanced class weighting, demonstrating its effectiveness for Malayalam-English code-mixed comments. DistilBERT achieved 0.68 accuracy on Task 2 (multiclass classification) with a learning rate of 5e-5, batch size of 8, weight decay of 0.01, and three epochs, indicating its ability to classify nuanced Malayalam news. However, the confusion matrix revealed a bias toward the False category, necessitating class weighting and enhanced preprocessing to correct the class imbalance. Figure 4 and figure 5 shows their classification reports respectively.

```
Classification Report:
              precision    recall  f1-score   support

        Fake       0.88      0.74      0.80       342
    original       0.75      0.89      0.82       310

    accuracy                           0.81       652
   macro avg       0.82      0.82      0.81       652
weighted avg       0.82      0.81      0.81       652
```

Figure 4: Classification Report of SVM with TF-IDF.

```
Classification Report:
              precision    recall  f1-score   support

       FALSE       0.68      1.00      0.81       260
   HALF TRUE       0.00      0.00      0.00        29
MOSTLY FALSE       0.00      0.00      0.00        23
PARTLY FALSE       0.00      0.00      0.00        58
 MOSTLY TRUE       0.00      0.00      0.00        10

    accuracy                           0.68       380
   macro avg       0.14      0.20      0.16       380
weighted avg       0.47      0.68      0.56       380
```

Figure 5: Classification Report of DistilBERT.

5 Conclusion

The purpose of this study was to detect fake news in Malayalam news articles and Malayalam-English code-mixed YouTube comments using various machine learning and deep learning algorithms. The study addressed issues such as code mixing, linguistic variances, and data scarcity while investigating successful categorization approaches. Our findings add to Dravidian language processing by comparing several ways to spotting disinformation. This Link contains the various algorithms used for this study. Future research can investigate data augmentation, multimodal techniques, and improved deep learning models to improve fake news identification.

6 Limitations

The results show that the model performs incorrectly when separating closely related classes in the multi-class classification problem, resulting in class overlap. Furthermore, while the binary classification performed well, it did occasionally misclassify borderline cases, demonstrating difficulties in dealing with subtle contextual distinctions.

References

R Anitha, S Navaneeth, Meharuniza Nazeem, and RR Rajeev. 2024. Code mixed english-malayalam sentiment analysis and sarcasm detection. In *2024*

15th International Conference on Computing Communication and Networking Technologies (ICCCNT), pages 1–6. IEEE.

Girma Bade, Olga Kolesnikova, Grigori Sidorov, and José Oropeza. 2024. Social media fake news classification using machine learning algorithm. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 24–29.

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.

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.

Raja Eduri, Soni Badal, Borgohain Samir Kumar, and Lalrempuii Candy. 2023. Dravidian fake news detection with gradient accumulation based transformer model. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 466–471.

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.

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.

S Kumar, M Anand Kumar, and KP Soman. 2019. Deep learning based part-of-speech tagging for malayalam twitter data (special issue: deep learning techniques for natural language processing). *Journal of Intelligent Systems*, 28(3):423–435.

M Madhumitha, M Kunguma, J Tejashri, et al. 2024. Techwhiz@ dravidianlangtech 2024: Fake news detection using deep learning models. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 200–204.

Aiswarya Mohan, Nafla Iqbal, and Manaal Mashpher. 2024. Malayalam fake news detection using optimized convolutional neural network (opcnn). In *2024 11th International Conference on Advances in Computing and Communications (ICACC)*, pages 1–6. IEEE.

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.

Rahul Ponnusamy, Kathiravan Pannerselvam, R Saranya, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, S Bhuvaneswari, Anshid Ka, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in tamil and malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488.

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.

Eduri Raja, Badal Soni, and Sami Kumar Borgohain. 2023. 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.

K Rameesa and KT Veeramanju. A systematic review on various approaches for news headlines categorization in malayalam language.

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.

Adhish S Sujan, Aleena Benny, VS Anoop, et al. 2023. Malfake: A multimodal fake news identification for malayalam using recurrent neural networks and vgg-16. *arXiv preprint arXiv:2310.18263*.

Habeeb Rahman YP and Muhammadali Nelliyullathil. 2023. Spread of misinformation in malayalam: A case study on the user engagement and impact of third-party fact-checking on facebook.