Beyond_Tech@DravidianLangTech2024 : Fake News Detection in Dravidian Languages Using Machine Learning

Kogilavani Shanmugavadivel[1], **Malliga Subramanian**[1], **Sanjai R**[1],
Mohammed Sameer B[1], **Motheeswaran K**[1]
[1]Department of AI, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{sanjair.22aid, mohammedsameerb.22aid}@kongu.edu
motheeswarank.22aid@kongu.edu

Abstract

In the digital age, identifying fake news is essential when fake information travels quickly via social media platforms. This project employs machine learning techniques, including Random Forest, Logistic Regression, and Decision Tree, to distinguish between real and fake news. With the rise of news consumption on social media, it becomes essential to authenticate information shared on platforms like YouTube comments. The research emphasizes the need to stop spreading harmful rumors and focuses on authenticating news articles. The proposed model utilizes machine learning and natural language processing, specifically Support Vector Machines, to aggregate and determine the authenticity of news. To address the challenges of detecting fake news in this paper, describe the Machine Learning (ML) models submitted to 'Fake News Detection in Dravidian Languages' at DravidianLangTech@EACL 2024 shared task. Four different models, namely: Naive Bayes, Support Vector Machine, Random forest, and Decision tree.

1 Introduction

People are increasingly choosing to search for and consume news from social media rather than traditional news sources as more and more of our lives are spent communicating online via social media platforms Albahr and Albahar (2020). Coelho et al. (2023) Fake news propagators have an opportunity to intentionally sway people's attitudes, beliefs, and trust by disseminating fake information. Rumors and false information typically travel quickly, harming specific relationships and social ties. Moreover, negative understanding, public scrutiny, and social distancing can also result in worry and emotional torment. Sharma et al. (2020) It is now necessary to relate to and filter out comparable false news automatically in order to lessen the harm and pain that fake news causes associations and communities. The internet and social media have made it much easier and more straightforward to obtain news information. It is true Gilda (2017) that there are a lot of websites that easily generate fake news. They usually use social media to boost their online presence and increase their impact. Dummy news websites pose as authoritative sources on topics (often political) in an attempt to sway public opinion. Jain et al. (2019) Fake information may be a global problem as well as a global task. Many experts think AI and machine literacy might potentially be used to address the problem of fake news. The paper is mainly concentrated on classifying whether a piece of news is fake or not.

In this paper, Problem and system description describes the dataset and how the dataset is preprocessed. The methodology uses classification algorithms to find the accuracy of models in classifying real and fake news in the given dataset and it also describes the algorithms. At last, the result gives the best model and its accuracy.

2 Literature Review

Ahmad and Lokeshkumar (2019) investigated text mining for the identification of false news. The dataset is initially preprocessed and relative algorithms are applied. Smitha and Bharath (2020) have taken data from many websites. The collected data are split into test and train and then the dataset is preprocessed, the preprocessed data are given to the ML algorithm after performing feature extraction.

The study Albahr and Albahar (2020) looks at random forests, Naive Bayes, and decision trees. The LIAR dataset, a popular dataset for identifying false news, was used for the experiment. To enhance the effectiveness of machine learning algorithms in identifying false news, They have employed NLP techniques. A variety of classification techniques, such as SVM, Bounded Decision Trees, Random Forests, Gradient Boosting, and Stochastic Gradient Descent, were employed byShaikh

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 124–128
March 22, 2024 ©2024 Association for Computational Linguistics

and Patil (2020). According to the Gilda (2017), TF-IDF of bi-grams fed into a Stochastic Gradient Descent model can identify non-credible sources with an accuracy of 77.2%.

Sharma et al. (2020) uses a machine learning classifier. After researching and using four distinct classifiers to train the model, They selected the most effective classifier for best model.

Jain et al. (2019) presented a method that combined SVM, and the Naive Bayes classifier. The three-part approach combines typical language preparation methods with machine learning calculations that split into controlled learning processes. In Coelho et al. (2023) they removed noise from the dataset containing Malayalam code-mixed data. They used ML models such as SVM and Random forest .Mandical et al. (2020) suggested to use hard voting with machine learning model as Multinomial Naive Bayes technique to detect bogus news in code-mixed Malayalam text.

Malliga et al. (2023) Shared task focused on categorizing social media posts in Malayalam using machine learning and transformer-based models. XLMRoBERTa-based model achieved exceptional performance with F1-score of 0.90.

3 Problem and System Description

Identifying and minimizing bogus news on social media is the aim of this collaborative effort on fake news identification.

3.1 Dataset Description

The shared task provides the dataset that is being utilized here. This project's main objective is to create machine learning-based model that can distinguish between authentic and bogus news.

Dataset	Original	Fake	Total
Training	1658	1,599	3,257
Testing	384	635	1,019

Table 1: Dataset Description

3.2 Preprocessing

The dataset consists of comments and their related labels such as fake and original. LabelEncoder is used to convert the categorical labels into numerical values as 0's and 1's.

4 Methodology

The methodology investigates a number of machine learning strategies and pre-processing techniques for the identification of fake news. The Naive Bayes classifier, SVM, Random Forest, and Decision Tree method are a few well-known classifiers that have been studied. The several steps taken while processing a text in order to classify it.

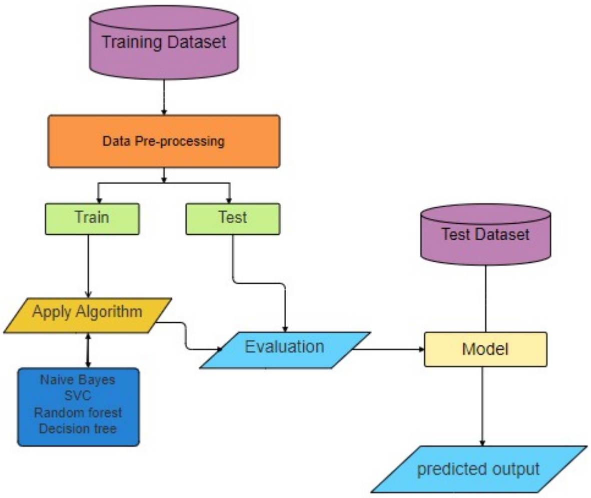
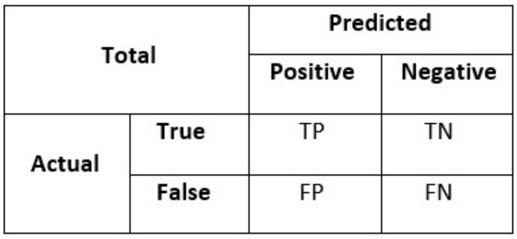


Figure 1: Processed workflow

4.1 Confusion Matrix

A confusion matrix serves as a tabular representation commonly employed to assess how well a classification model performs on a given set of test data with known true values. This matrix facilitates a visual depiction of the algorithm's performance, offering insights into its accuracy and error patterns.



Total		Predicted	
		Positive	Negative
Actual	True	TP	TN
	False	FP	FN

Figure 2: Confusion matrix

True Positive (TP) occurs when the model correctly identifies fake news as fake.
True Negative (TN) occurs when the model correctly classifies true news as true.

False Negative (FN) happens when the model mistakenly categorizes true news as fake.
False Positive (FP) happens when the model incorrectly labels fake news as true.

4.2 Naive Bayes classifier

Naive Bayes is a probabilistic algorithm that assumes features are independent for quick decision-making. It's often used in text classification and spam filtering, making predictions based on simple assumptions.

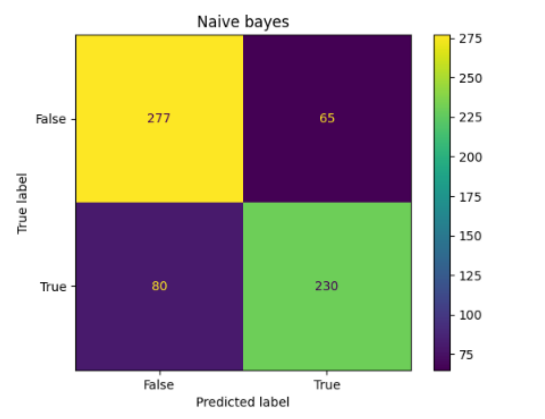
This can be stated as:

$$P(Y|X_1, X_{2,...}X_n) = \frac{P(X_1|Y)P(X_2|Y) \ldots P(X_n|Y)}{P(X_1)P(X_2) \ldots P(X_n)}$$

which can be further expressed as:

$$P(Y|X_1, X_{2,...}X_n) = \frac{P(Y)\prod_{i=1}^{n} P(X_i|Y)}{P(X_1)P(X_2) \ldots P(X_n)}$$

where P(X—Y) is the likelihood that event X will occur given that event Y has already occurred.



```
Accuracy: 0.7776073619631901
```

Figure 3: Confusion Matrix for Navie Bayes

	Precision	Recall	f1-score
Accuracy			0.78
Macro avg	0.78	0.78	0.78
Weighted avg	0.78	0.78	0.78

Table 2: Classification Report for Naive Bayes classifier

4.3 Support Vector Machine

SVM is an effective machine learning method for regression and classification that divides data classes into groups by finding the best hyperplane in high-dimensional space. support vectors are used to establish the decision boundary, SVM is resistant to overfitting. For big datasets, SVM can be computationally demanding despite its efficacy.

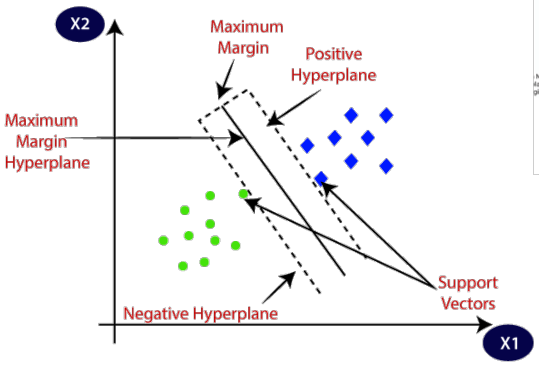
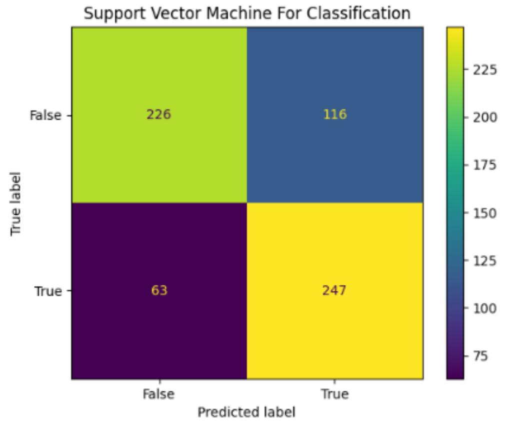


Figure 4: Support vector Machine Graph



```
Accuracy: 0.7776073619631901
```

Figure 5: Confusion Matrix for SVM

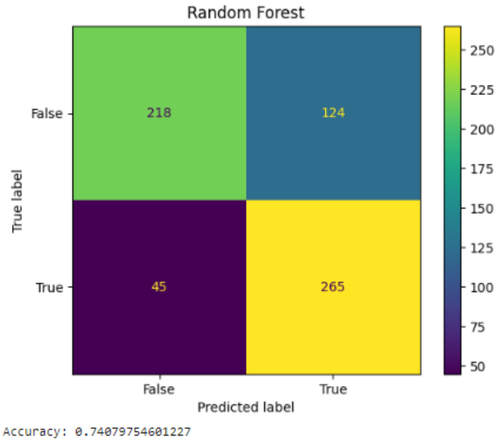
	Precision	Recall	f1-score
Accuracy			0.73
Macro avg	0.73	0.73	0.73
Weighted avg	0.73	0.73	0.72

Table 3: Classification Report for SVM

4.4 Random Forest

Random Forest in machine learning is like a diverse group of decision-making experts collaborating on a complex problem. It constructs multiple decision trees, each with its perspective on the data.

Individually, these trees may have limitations, but collectively, they form a robust and versatile ensemble. The forest's strength lies in aggregating these diverse insights, reducing overfitting, and delivering a more accurate and reliable prediction, making it a go-to choice for various tasks, from classification to regression.



```
Accuracy: 0.74079754601227
```

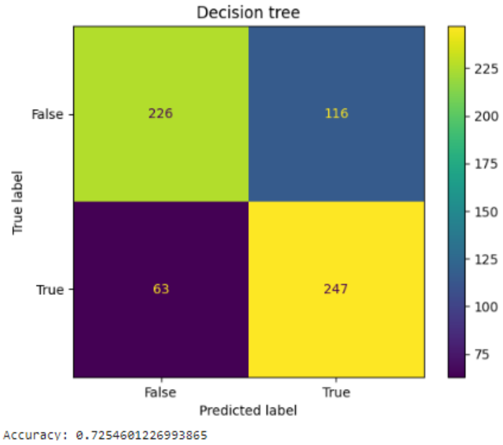
Figure 6: Confusion Matrix for Random Forest

	Precision	Recall	f1-score
Accuracy			0.74
Macro avg	0.76	0.75	0.74
Weighted avg	0.76	0.74	0.74

Table 4: Classification Report for Random Forest

4.5 Decision Tree

A decision tree, in supervised learning, structures attribute tests in a tree-like form for classification and regression. Nodes represent tests, branches show outcomes, and leaf nodes hold class labels. Attributes are chosen during training using metrics like entropy or Gini impurity for optimal information gain. The decision tree is recursively built, starting from the root node, until meeting stopping criteria like maximum depth. Impurity measures, such as Gini index or entropy, assess randomness, while pruning removes non-informative branches to prevent overfitting.



```
Accuracy: 0.7254601226993865
```

Figure 7: Confusion Matrix for Decision Tree

	Precision	Recall	f1-score
Accuracy			0.73
Macro avg	0.73	0.73	0.73
Weighted avg	0.73	0.73	0.72

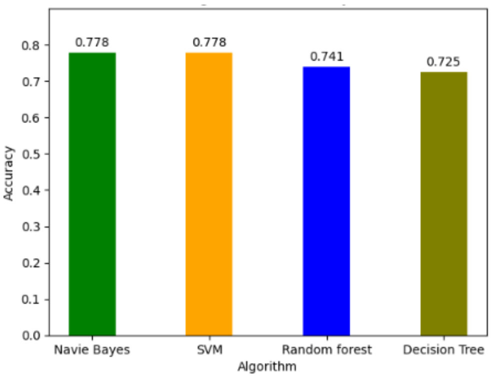
Table 5: Classification Report for Decision Tree

5 Result

A good dataset is first used to train the model. Second, many performance metrics are used to evaluate performance. Lastly, headlines or articles are categorized using the best model—that is, the model with the highest accuracy. At 77.7%, Navie bayes and SVM proved to be the most effective model for static search.

6 Conclusion

Finally, it should be noted that when fake news spreads, it attempts to alter people's perceptions and attitudes about utilizing digital technologies. There are two possible outcomes when individuals fall for fake news: Initially, people begin to think that their preconceived notions about a given subject are accurate. Our fraudulent News Detection System was developed to stop this problem by evaluating user-submitted information and classifying it as real or fraudulent. Several machine learning and natural language processing (NLP) approaches must be used to do this.



Classifier	Accuracy
Naive Bayes classifier	77.76
Support Vector Machine	77.76
Random Forest	74.07
Decision Tree	72.54

Table 6: algorithm and accuracy

References

Faraz Ahmad and R Lokeshkumar. 2019. A comparison of machine learning algorithms in fake news detection. *International Journal on Emerging Technologies*, 10(4):177–183.

Abdulaziz Albahr and Marwan Albahar. 2020. An empirical comparison of fake news detection using different machine learning algorithms. *International Journal of Advanced Computer Science and Applications*, 11(9).

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.

Shlok Gilda. 2017. Notice of violation of ieee publication principles: Evaluating machine learning algorithms for fake news detection. In *2017 IEEE 15th student conference on research and development (SCOReD)*, pages 110–115. IEEE.

Anjali Jain, Avinash Shakya, Harsh Khatter, and Amit Kumar Gupta. 2019. A smart system for fake news detection using machine learning. In *2019 International conference on issues and challenges in intelligent computing techniques (ICICT)*, volume 1, pages 1–4. IEEE.

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.

Rahul R Mandical, N Mamatha, N Shivakumar, R Monica, and AN Krishna. 2020. Identification of fake news using machine learning. In *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pages 1–6. IEEE.

Jasmine Shaikh and Rupali Patil. 2020. Fake news detection using machine learning. In *2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)*, pages 1–5. IEEE.

Uma Sharma, Sidarth Saran, and Shankar M Patil. 2020. Fake news detection using machine learning algorithms. *International Journal of Creative Research Thoughts (IJCRT)*, 8(6):509–518.

N Smitha and R Bharath. 2020. Performance comparison of machine learning classifiers for fake news detection. In *2020 Second international conference on inventive research in computing applications (ICIRCA)*, pages 696–700. IEEE.