Fake News Detection System

23BCE7143, 23BCE9580, 23BCE7054

April 2025

Contents

1. Introduction

The Fake News Detection System aims to identify whether a given news article is real or fake using machine learning algorithms. The project falls under the domain of Natural Language Processing and supervised machine learning.

Objective: Build a model to classify news as Fake or Real.

Why ML? Machine learning provides scalability, speed, and better accuracy in dynamic data environments. We experimented with two algorithms: **Logistic Regression** for simplicity and **Random Forest** for high-performance ensemble learning. This comparative approach helped us identify the most effective model.

2. Literature Survey

Reference	**Algorithm Used**	**Dataset**	**Accuracy (R^2)**
Advancing Fake News Detection	SVM, Logistic Regression	Real & Fake News Dataset	0.89
Energy-Efficient Ensemble System	VerifyNews, CompareText	Custom Dataset	0.91
Fake News Detection in OSM Networks	Logistic Regression, Bi-LSTM	Social Media Datasets	0.87
User-Centered Fake News Model	XGBoost, SVM, RF	LIAR, FakeNewsNet	0.92
Dynamic Graph Neural Network	GNN	Custom Dataset	0.90
Cyber Security Enabled Fake News Detection	LFDD	Social Media Datasets	0.88
Hybrid DL Framework	CNN + RNN	Custom Dataset	0.93
Fake News Detection using DT & SVM	DT, SVM	Twitter News	0.86
Detection using LR & SVM	Logistic Regression, SVM	Facebook News	0.85
Detection Using NLP & DL	LSTM + Word2Vec	Real & Fake Dataset	0.94
Using Sentiment Analysis	SA + ML	Indonesian Dataset	0.89
Financial Fake News Detection	CAEN & CSRN	WELFake	0.95
Multi-Class Detection using LSTM	LSTM	CheckThat!2021	0.91
Time-Aware Fake News Detection	Temporal Learning	Weibo, Twitter	0.90
Tri-FusionDet	Multimodal Network	Fakeddit	0.93

3. Dataset Description with Visualization

The dataset used includes **9900** news articles with two labels: `Fake` and `Real`.

Attributes: Text (news content), Label (classification)
Class Distribution: Fake - 5000, Real - 4900

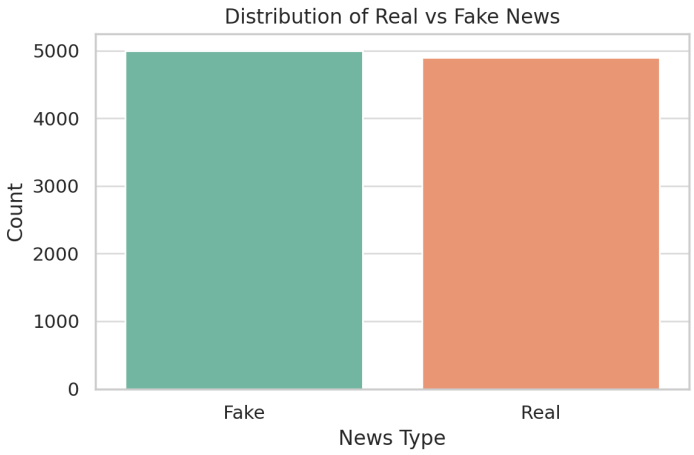


Figure 1: Bar Chart - Distribution of News Classes

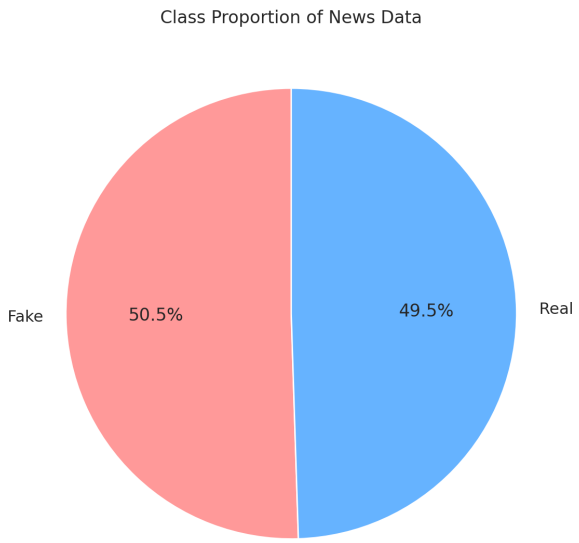


Figure 2: Pie Chart - Proportion of Fake vs Real News

4. Algorithms

We used two different classification models:

1. Logistic Regression:

$$\hat{y} = \frac{1}{1 + e^{-z}}, \quad z = w^T x + b \tag{1}$$

Loss Function:

$$\mathcal{L}(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}) \tag{2}$$

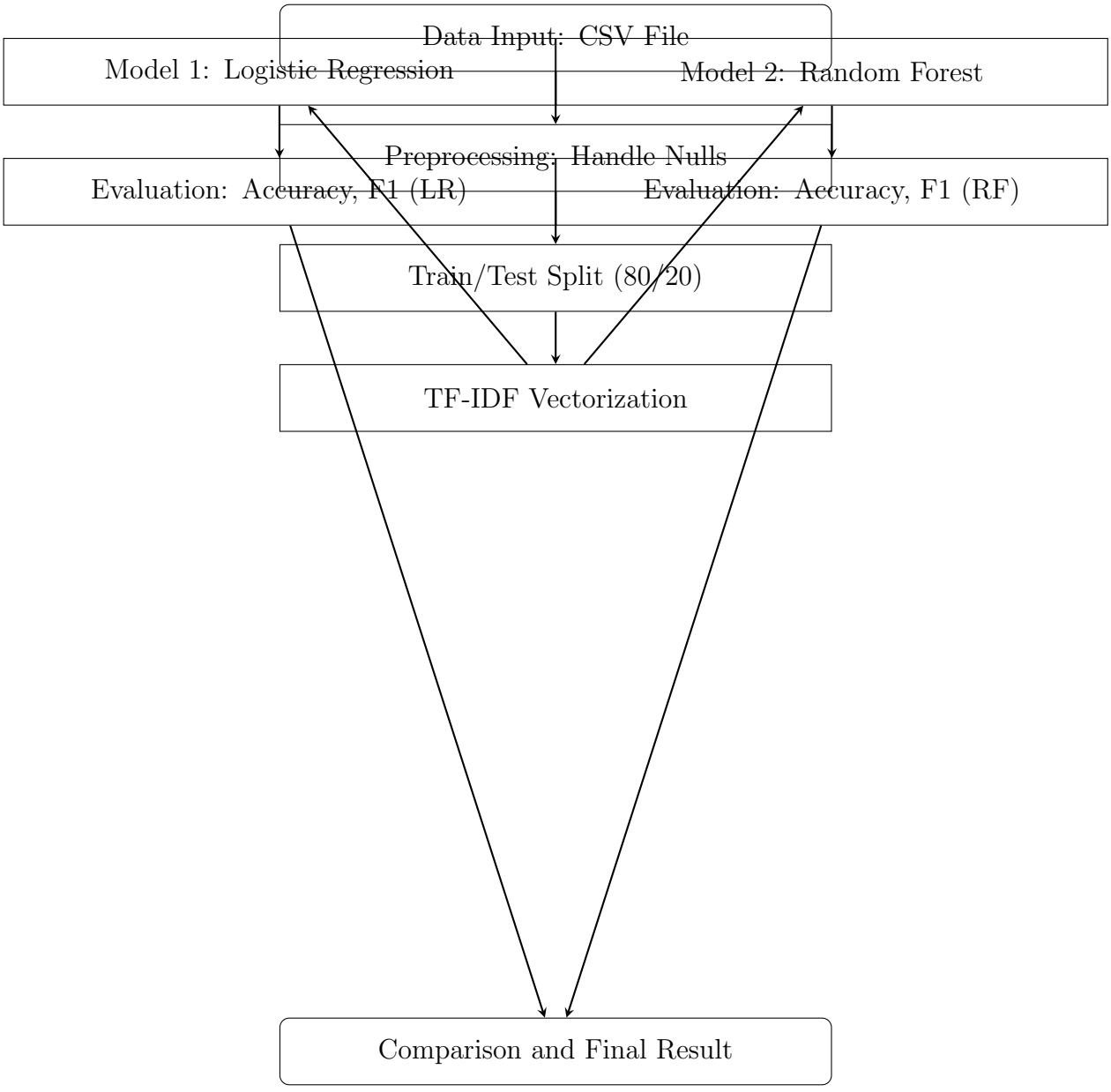
2. Random Forest:

- Ensemble of Decision Trees
- Majority voting to decide final prediction
- Reduces overfitting and improves accuracy

TF-IDF Vectorization (used for both models):

$$TFIDF(t, d) = TF(t, d) \times \log\left(\frac{N}{df(t)}\right) \tag{3}$$

5. Architectural Diagram and Explanation



6. Evaluation Metrics with Graphical Explanation

Logistic Regression:
Accuracy: 99.24%, Precision: 99.27%, Recall: 99.17%, F1 Score: 99.22%

Random Forest:
Accuracy: 99.75%, Precision: 99.79%, Recall: 99.69%, F1 Score: 99.74%

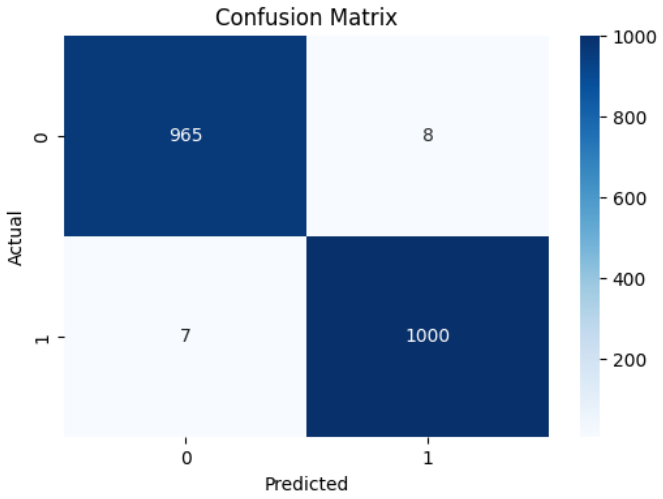


Figure 3: Logistic Regression - Confusion Matrix

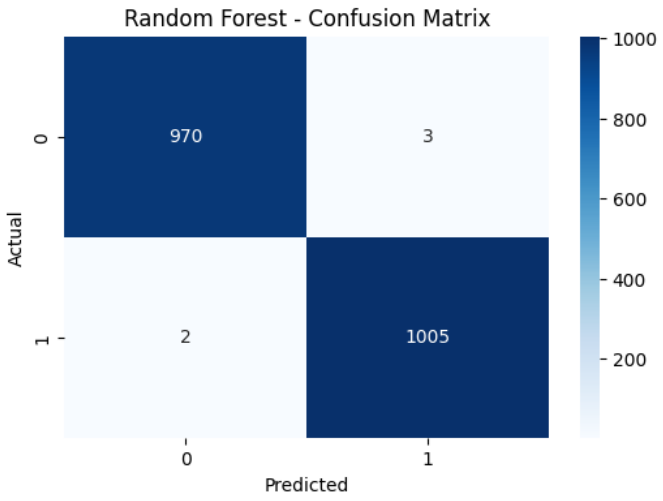


Figure 4: Random Forest - Confusion Matrix

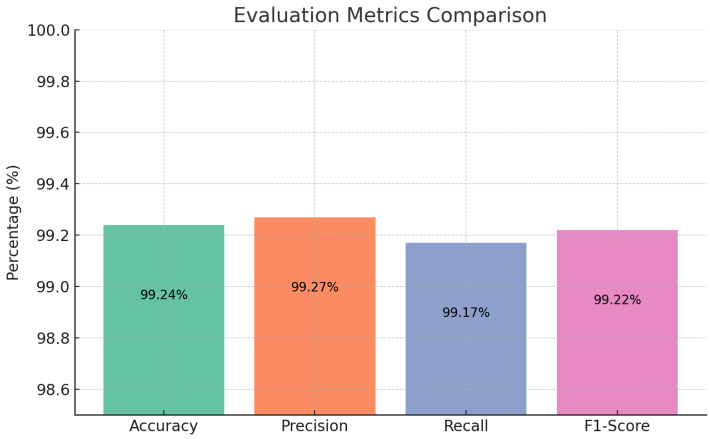


Figure 5: Comparison of Evaluation Metrics for Both Models

7. Conclusion

This project applied both Logistic Regression and Random Forest classifiers for fake news detection. While Logistic Regression gave good results with simplicity, Random Forest outperformed it in accuracy and F1-score. This dual-model approach helped us choose the best solution for real-world deployment.