Dayananda Sagar University

School of Engineering

Devarakaggalahalli, Harohalli, Kanakapura Road, Ramanagara Dt., Bengaluru – 562 112

Department of
Computer Science & Engineering

Mini Project Report
on

Artificial Intelligence & Machine Learning

FAKE NEWS DETECTION

By

K Kavya - ENG22CS0076
Kotigi Jyothi - ENG22CS0082
Kruthika B N - ENG22CS0084

Under the supervision of

Mrs. Shilpa Sudheendran

Assistant Professor,
Department of Computer Science and Engineering

Dayananda Sagar University

School of Engineering

Devarakaggalahalli, Harohalli, Kanakapura Road, Ramanagara Dt., Bengaluru – 562 112

Department of Computer Science & Engineering

CERTIFICATE

This is to certify that the work titled **“FAKE NEWS DETECTION ”** is carried out by **K Kavya(ENG22CS0076), Kotigi Jyothi(ENG22CS0082), Kruthika B N (ENG22CS0084)** Bonafide students of Bachelor of Technology in Computer Science and Engineering at the School of Engineering, Dayananda Sagar University, Bangalore in partial fulfillment for the award of degree in Bachelor of Technology in Computer Science and e, during the year **2024-2025**.

Mrs. Shilpa Sudheendran	**Dr. Girisha G S**
Assistant Professor, Dept. of CSE,	Chairperson CSE,
School of Engineering,	School of Engineering,
Dayananda Sagar University.	Dayananda Sagar University.

Dayananda Sagar University

School of Engineering

Devarakaggalahalli, Harohalli, Kanakapura Road, Ramanagara Dt., Bengaluru – 562 112

Department of Computer Science & Engineering

DECLARATION

We, **K Kavya(ENG22CS0076), Kotigi Jyothi(ENG22CS0082), Kruthika B N (ENG22CS0084)**are students of the fifth semester B.Tech in Computer Science and Engineering, at School of Engineering, Dayananda Sagar University, hereby declare that the mini-project titled "**FAKE NEWS DETECTION** " has been carried out by us and submitted in partial fulfillment for the award of degree in Bachelor of Technology in Computer Science and Engineering during the academic year 2024-2025.

Student Signature
Name1: K Kavya
USN : ENG22CS0076
Name2: Kotigi Jyothi
USN : ENG22CS0082
Name3: Kruthika B N
USN : ENG22CS0084

Place : Bangalore

Date :

ABSTRACT

Fake news poses significant challenges to the credibility of information in the digital age. This study focuses on leveraging Machine Learning techniques, specifically Logistic Regression and Random Forest algorithms, for detecting fake news. The dataset used includes labeled news articles, categorized as either "fake" or "real," processed through text preprocessing techniques such as tokenization, stop-word removal, and TF-IDF vectorization.

Logistic Regression, a probabilistic linear model, offers simplicity and interpretability in analyzing textual data for classification tasks. Conversely, the Random Forest, a non-linear model, provides a robust mechanism for handling complex relationships within the data. Both models are trained and evaluated on key performance metrics, including accuracy, precision, recall, and F1-score.

The results indicate that while Logistic Regression demonstrates consistent performance due to its reliance on linear decision boundaries, Random Forest models excel in capturing non-linear patterns but may be prone to overfitting. A comparative analysis underscores the strengths and limitations of these approaches, providing insights into their applicability for fake news detection. Future work involves exploring ensemble methods and deep learning techniques to enhance classification performance.

INTRODUCTION

Fake news has become a pervasive issue in the digital era, fueled by the rapid growth of social media and online platforms. Misinformation can influence public opinion, cause social unrest, and have far-reaching consequences on society. Detecting fake news is a crucial challenge in the field of Artificial Intelligence and Machine Learning (AI/ML), as it involves analyzing textual data to distinguish between authentic and fabricated content.

This project focuses on implementing a fake news detection system using Logistic Regression and Randon Forest algorithm. These are widely used machine learning techniques that offer complementary approaches to solving classification problems. Logistic Regression provides a probabilistic framework for binary classification, while Randon Forest use multiple decision tree for decision-making based on feature splits.

The proposed system aims to classify news articles as either fake or real by leveraging features extracted from their textual content. The process involves data preprocessing, feature engineering, model training, and evaluation. By comparing the performance of Logistic Regression and Randon Forest classifiers, the project seeks to identify their respective strengths and weaknesses in the context of fake news detection.

PROJECT DESCRIPTION

The **Fake News Detection** project focuses on developing a robust machine learning model to predict fake or real news based on various features influencing airfare. The project entails the following key steps and highlights:

1. **Objective**

   The objective of this project is to develop a machine learning system capable of accurately detecting fake news by classifying news articles as either fake or real. The project focuses on preprocessing textual data, implementing Logistic Regression and Random Forest classifiers, and evaluating their performance. Additionally, it aims to analyze the strengths and weaknesses of these algorithms to determine their suitability for the task.

2. **Dataset**

   The project uses a labeled dataset containing news articles categorized as fake or real. The dataset, often sourced from platforms like Kaggle, includes thousands of records with features such as headlines, full articles, and labels indicating authenticity. The data format is typically CSV or JSON, with sources from both legitimate and fabricated news outlets.

3. **EDA**

   Exploratory Data Analysis is performed to understand the dataset's structure and distribution. This involves analyzing the proportion of fake versus real articles, variations in text length, word frequencies, and addressing any missing data. Insights from this analysis guide subsequent preprocessing and feature engineering steps.

4. **Feature Engineering**

   The project converts textual data into numerical formats for machine-readability. Text preprocessing includes removing punctuation, stopwords, and applying stemming or lemmatization. Feature extraction methods like TF-IDF, Bag of Words, and N-grams are used to effectively represent the text. Dimensionality reduction techniques are applied to optimize the feature space, enhancing model performance.

5. Model Development

The project employs two machine learning models:

- **Logistic Regression:** A probabilistic linear model well-suited for binary classification tasks.
- **Random Forest:** An ensemble method based on decision trees, which improves classification accuracy by combining the predictions of multiple trees.

Both models are fine-tuned through hyperparameter optimization and validated using cross-validation techniques to ensure robustness and accuracy.

6.Results

The models are evaluated using metrics such as accuracy, precision, recall, and F1-score. Logistic Regression typically demonstrates strong performance on datasets with linear relationships. Random Forest excels at capturing complex patterns in the data and reduces overfitting compared to individual decision trees. A confusion matrix is used to provide a detailed breakdown of the classification results, including true positives, true negatives, false positives, and false negatives.

7.Impact

This fake news detection system addresses the growing issue of misinformation by providing an automated and efficient solution for analyzing and classifying news articles. Its societal impact includes curbing the spread of false information and fostering informed decision-making. On a technological level, it demonstrates the practical application of machine learning in solving real-world challenges.

8.Conclusion

The project successfully applies Logistic Regression and Random Forest algorithms to detect fake news, showcasing their utility in text classification tasks. Random Forest, in particular, offers robust performance for non-linear data, complementing Logistic Regression's strengths.

Social and Environmental Impact of the FAKE NEWS DETECTION

Social Impact

Positive Impacts

1. **Improved Public Awareness and Critical Thinking**
   - Fake news detection tools help educate the public on identifying misinformation, fostering critical thinking and informed decision-making.
2. **Reduction in Harmful Misinformation**
   - By identifying and mitigating fake news, these tools reduce the spread of harmful or divisive content, contributing to a healthier public discourse.
3. **Enhanced Trust in Media**
   - Reliable fake news detection can restore trust in credible media outlets and platforms by ensuring accurate information dissemination.
4. **Support for Democracy**
   - Combatting fake news during elections or political campaigns helps preserve democratic processes by curbing voter manipulation.

Negative Impacts

1. **Censorship Concerns**
   - Automated tools may mistakenly flag legitimate content as fake, potentially infringing on free speech and suppressing diverse viewpoints.
2. **Polarization Risks**
   - If perceived as biased, fake news detection tools could deepen societal divisions and reduce trust in institutions employing them.

3. **Unintended Biases**
    - Algorithms may reflect or amplify biases present in the training data, disproportionately affecting certain groups or viewpoints.
4. **Erosion of Privacy**
    - Advanced fake news detection often relies on data analysis, raising concerns about the collection and use of personal information.

Environmental Impact

Positive Impacts

1.Reduction in Resource Waste

By reducing the proliferation of fake news, resources spent on debunking misinformation and addressing its consequences are minimized.

Negative Impacts

1. **Energy Consumption**
    - Training and deploying machine learning models for fake news detection consume significant energy, contributing to carbon emissions. For example, large-scale AI models have a considerable carbon footprint.
2. **E-Waste Generation**
    - Hardware used in developing and running fake news detection systems (e.g., servers and GPUs) has a finite lifecycle, contributing to electronic waste.

Dataset Description

Dataset Overview

Title: Fake News Detection

Text: The main body of the news article.

Subject: The category or topic of the news (e.g., News, Politics).

Features

Title: A short text string summarizing the content of the article.

Text: A longer text field containing the detailed content of the article.

Subject: A categorical feature indicating the topic of the article, which could provide additional context for classification.

Date: A temporal feature that can indicate trends or patterns in fake news over time.

Statistical Insights

Total Records: 23,481 articles.

Non-null Values: All columns are complete, with no missing values.

Feature Types:

All features are of type object (text or categorical).

Subjects:

The "subject" column contains different topics/categories, which can provide valuable insights during analysis.

Data Preprocessing:

Cleaning the text and title columns by removing punctuation, special characters, and stopwords.
Converting text to lowercase for uniformity.

Date Parsing:
Converting the date column into a datetime format to analyze temporal trends.

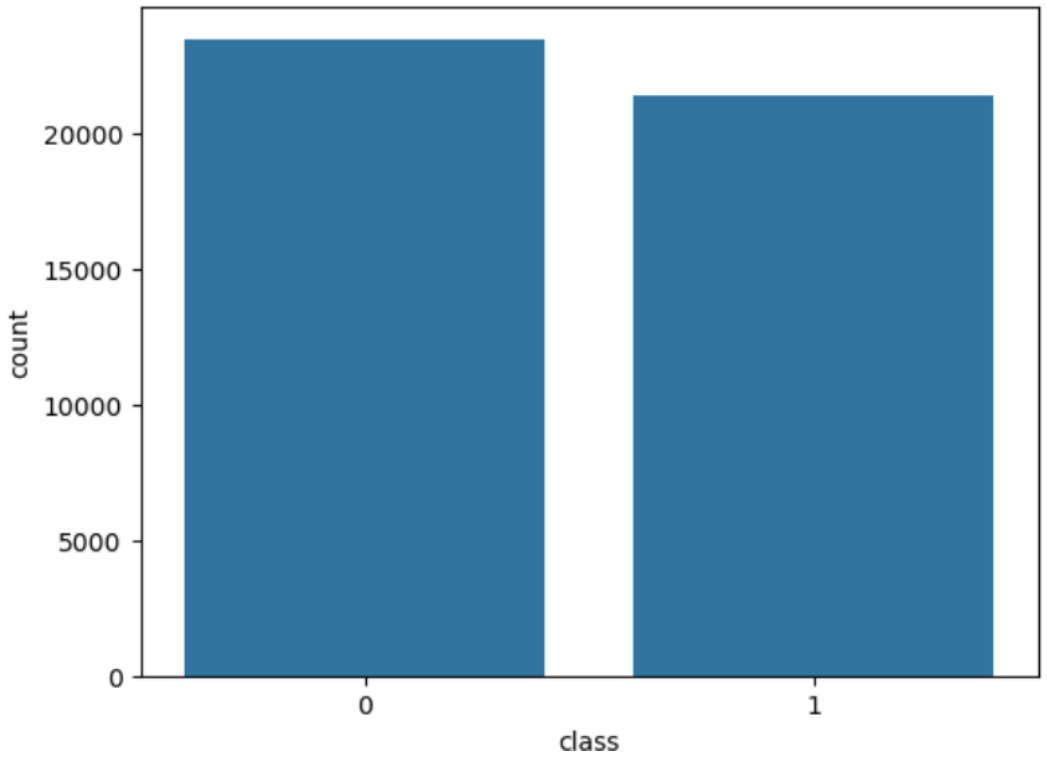
Feature Engineering:
Extracting features like word count, character count, and n-grams for model input.
Using techniques like TF-IDF or Bag of Words for vectorizing text data.

Categorical Encoding:
Encoding the "subject" column into numerical values if used in the model.

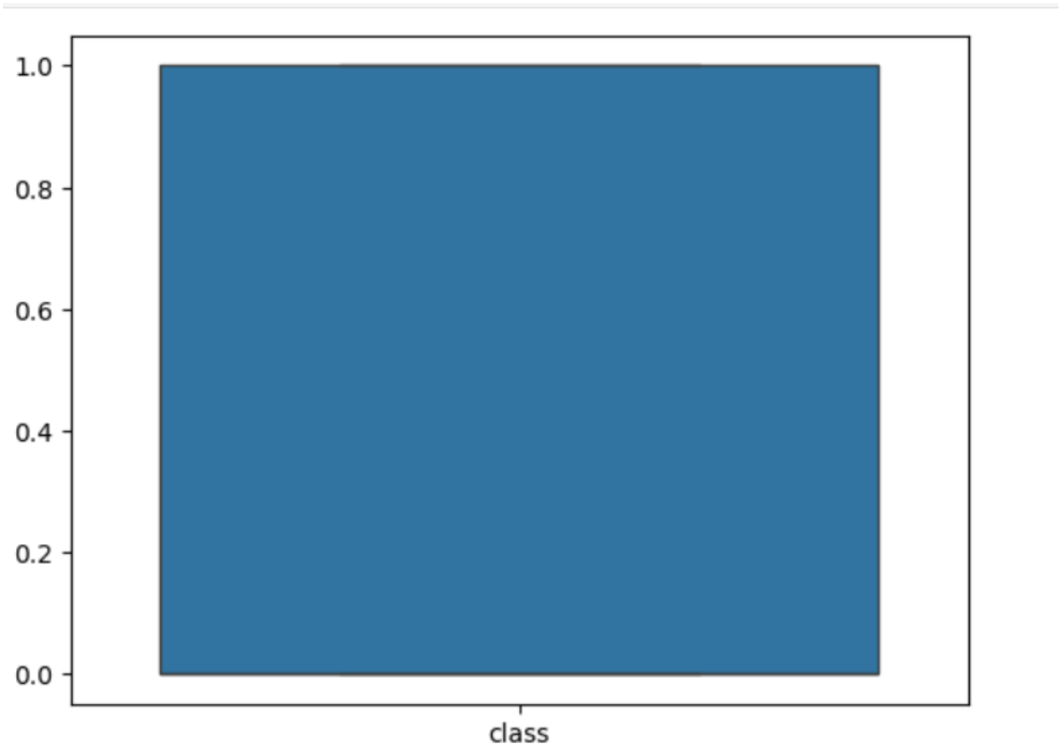
Train-Test Split:
Dividing the data into training and testing sets for model evaluation.

Exploratory Data Analysis (EDA)

EDA was conducted to gain insights into the factors influencing the classification of news articles as fake or real and to identify patterns in the dataset. Key relationships were explored between the target variable (Label) and various features such as text length, word frequency, and subject categories using visualizations and statistical summaries.



The bar chart shows a balanced distribution of two classes in a dataset, with nearly equal counts for class 0 (e.g., "fake news") and class 1 (e.g., "real news"). This balance is important for training Machine Learning models like Logistic Regression and Random Forest, as it ensures fair learning without bias toward one class. Balanced datasets generally lead to more reliable and meaningful evaluation metrics, such as accuracy and F1-score.



The chart shows that the dataset contains only one class, indicating no variability in the target variable. This lack of class diversity makes it unsuitable for classification tasks, as the model cannot learn to differentiate between multiple classes. Additionally, the chart does not provide information about outliers; other methods, such as analyzing feature distributions or using statistical techniques, would be needed to identify anomalies in the data. For our data set we don’t have the outliers.

Data Preprocessing and Model Preparation

1. **Label Encoding**:

- Converts categorical target labels (e.g., "fake" and "real") into numerical values (e.g., 0 and 1) to make them compatible with machine learning algorithms.
- Ensures consistent and machine-readable representation of the target variable.

2. Feature Selection:

- Identifies and retains the most important features (e.g., words or phrases) that contribute significantly to distinguishing between fake and real news.
- Reduces dimensionality, improves model performance, and minimizes noise. Methods like TF-IDF, Chi-Square tests, or Mutual Information can be used for this purpose.

3. Data Splitting:

- Splits the dataset into training and testing subsets to evaluate model performance.
- A typical split involves allocating 70-80% of the data for training and 20-30% for testing. This ensures that the model generalizes well on unseen data.

4. Feature Scaling:

- Standardizes or normalizes feature values to ensure uniform contribution to the model.
- While textual data processed through TF-IDF may not require scaling, additional numerical features (if present) should be scaled using techniques like standardization or normalization.

Training the Models

Logistic Regression is a widely used classification algorithm, particularly suited for binary classification tasks like fake news detection. In this case, it predicts whether a news article is fake or real based on features extracted from the text (e.g., word frequencies, TF-IDF scores). The model uses a logistic function to estimate probabilities between 0 and 1, which are then converted into class labels (fake or real). Logistic Regression is particularly effective when the relationship between input features and target labels is linear.

Random Forest is an ensemble learning method that combines multiple Decision Trees to improve the accuracy and robustness of predictions. For fake news detection, it predicts whether a news article is fake or real by aggregating the outputs of several Decision Trees, each trained on a random subset of the data and features. This approach reduces overfitting and increases generalization. Random Forests are powerful for datasets with complex, non-linear relationships and are less sensitive to noise.

Model Evaluation

After training the models, they were evaluated on the test set using various performance metrics, including:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- R^2 Score
- Adjusted R^2
- Mean Absolute Percentage Error (MAPE)
- Root Mean Squared Log Error (RMSLE)

Results

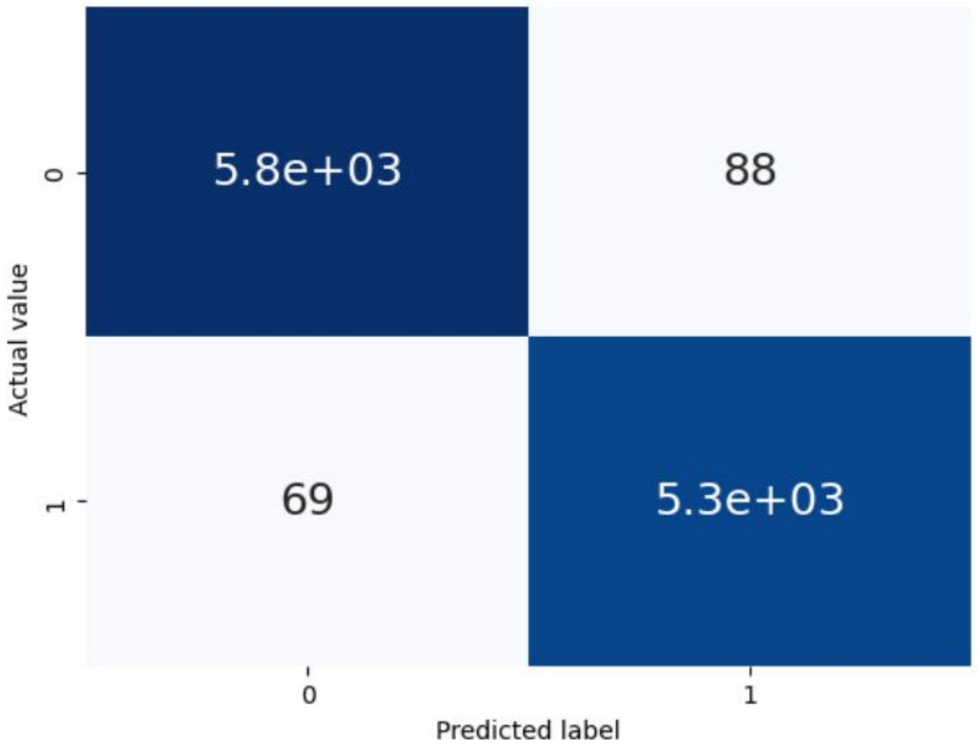
The results of a fake news detection model typically include the model's ability to classify news articles as fake or real, which can be evaluated using several metrics Accuracy: The percentage of correctly classified news articles (both fake and real) out of the total number of articles. A high accuracy indicates the model is correctly distinguishing between fake and real news.

Precision: Measures the proportion of correctly identified fake news articles out of all the articles the model labeled as fake. This is important if false positives (incorrectly labeling real news as fake) are costly.

Recall: Measures the proportion of actual fake news articles that were correctly identified by the model. This is critical when it's important to detect as many fake news articles as possible.

F1-Score: The harmonic mean of precision and recall, providing a balanced measure when both false positives and false negatives need to be minimized.

Confusion Matrix: A table that shows the number of true positives, true negatives, false positives, and false negatives, helping to visualize the model's performance.



The image shows a confusion matrix, which is commonly used in machine learning to evaluate the performance of a classification model. Here's what each part of the confusion matrix represents

Values in the Matrix

1. Top-left (5.8e+03 or 5800):

True Negatives (TN): The model correctly predicted 0 (negative class).

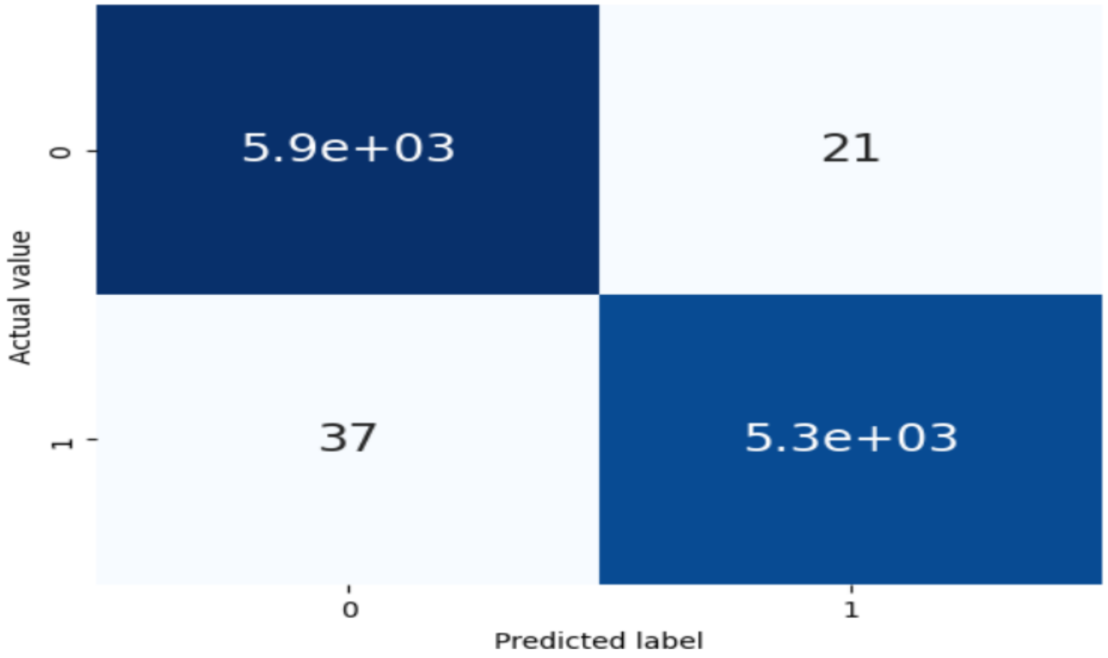
2. Top-right (88):

False Positives (FP): The model incorrectly predicted 1 (positive class) when the true label was 0.

3. Bottom-left (69):

False Negatives (FN): The model incorrectly predicted 0 (negative class) when the true label was 1.

1. Bottom-right (5.3e+03 or 5300): True Positives (TP): The model correctly predicted 1 (positive class).



The image shows a confusion matrix, a common evaluation tool used in machine learning to measure the performance of a classification model.

Confusion Matrix:

1. Axes:

Actual Value (Y-axis): Represents the true labels (ground truth) of the data.

Predicted Label (X-axis): Represents the labels predicted by the model.

2. Cells:

Top-left (True Negative - TN): 5900 (5.9e+03) instances were correctly classified as class 0.

Top-right (False Positive - FP): 21 instances were incorrectly classified as class 1 when they actually belonged to class 0.

Bottom-left (False Negative - FN): 37 instances were incorrectly classified as class 0 when they actually belonged to class 1.

Bottom-right (True Positive - TP): 5300 (5.3e+03) instances were correctly Classified as class 1.

`

Conclusion

On the other hand, the Random Forest model is more flexible and capable of capturing complex, non-linear relationships in the data. It constructs an ensemble of decision trees, where each tree is trained on a random subset of the data and features. By aggregating the predictions of multiple trees, Random Forest achieves higher accuracy and robustness, reducing the risk of overfitting compared to individual decision trees. This makes it particularly useful for distinguishing subtle patterns in fake news articles, even when the dataset is noisy or the feature interactions are intricate.

In practice, both models can serve as a good starting point for fake news detection. Logistic Regression is often the go-to choice for a quick, interpretable baseline. It performs well when the relationship between features (e.g., word frequencies) and the target label (fake or real) is linear. Logistic Regression provides fast training, easy interpretability, and is effective for high-dimensional text data. However, it may struggle to capture complex, non-linear patterns.

On the other hand, Random Forest excels in scenarios where feature interactions and non-linear relationships are critical to the task. Depending on the dataset's characteristics and performance requirements, these models can be further optimized or combined with ensemble methods for better accuracy and generalization. In the task of fake news detection, both Logistic Regression and Random Forest offer valuable approaches, each with its strengths and weaknesses