Fake News Detection and Evaluation

Sudhanshu Kumar

Internship in Data Science (Fake News Detection Project)

B.Tech in IT and B.P.Poddar Institute of Management and Technology

Period of Internship: 25th August 2025 – 19th September 2025

Report submitted to: IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata

● Abstract

This project is about detecting fake news using machine learning. Nowadays, fake news spreads very fast on social media, so we wanted to build a model that can automatically classify whether a news article is real or fake. For this project, we worked with datasets of real and fake news. We cleaned the text by removing unnecessary words and characters, then converted it into numbers using methods like Word2Vec and TF-IDF. After that, we trained different models such as Logistic Regression, Random Forest, and AdaBoost. We compared their performance and found that TF-IDF combined with AdaBoost gave the best accuracy. The results showed that machine learning can help in identifying fake news with good reliability.

● Introduction

Fake news is a serious problem today because it can mislead people and create confusion. This project is relevant as it helps in automatically detecting fake news articles using data science and machine learning techniques.

The technology involved in this project includes **Python, Natural Language Processing (NLP), and Machine Learning algorithms**. We studied background materials such as research papers on fake news detection and tutorials on NLP.

The purpose of doing this project was to learn how text data can be processed and how machine learning models can be applied for classification tasks.

Topics learned during the first two weeks of training:

- Basics of Python and Jupyter Notebook
- Handling datasets with Pandas and NumPy
- Visualization using Matplotlib and Seaborn
- Basics of Machine Learning and model evaluation
- Introduction to Natural Language Processing (NLP)
- Text preprocessing (cleaning, tokenization, stopwords removal)
- Feature extraction techniques like Word2Vec and TF-IDF

• Project Objective

The main objectives of this project are:

- To understand the dataset of fake and real news.
- To clean and preprocess text data for machine learning.
- To apply feature extraction methods (Word2Vec and TF-IDF).
- To train machine learning models such as Logistic Regression, Random Forest, and AdaBoost.
- To compare the performance of models and identify which method works best for fake news detection.

• Methodology

In this project, our main goal was to develop a machine learning model to detect fake news from textual data. The work involved several steps, starting from data collection to model evaluation.

1. Data Collection:

- We collected a combined dataset of fake and true news articles from publicly available sources.
- The dataset contained columns such as `title`, `text`, `label` (0 = True, 1 = Fake), and `date`.
- Total number of articles: 44,898 (Fake: 23,481, True: 21,417).
- No survey was conducted in this project as we used secondary datasets.

2. Data Cleaning and Preprocessing:

- Removed punctuation, special characters, numbers, and URLs from the text.
- Converted all text to lowercase to maintain consistency.
- Removed English stop words to reduce noise in the data.
- Extra whitespace and newline characters were removed.
- Preprocessed text was stored in a new column `text_clean`.

3. Feature Extraction:

- TF-IDF vectorization was applied to convert text into numerical features for machine learning models.
- Parameters used: maximum 5000 features, ngram range = (1,2), minimum document frequency = 5, stop words = English.
- TF-IDF helped in giving weight to important words while ignoring common words.

4. Model Development:

- Two machine learning models were developed and compared:
    - **Random Forest Classifier:** Ensemble of multiple decision trees to reduce overfitting and improve accuracy.
    - **AdaBoost Classifier:** Boosting method that combines weak learners (decision stumps) to form a strong learner.
- Models were trained using the preprocessed TF-IDF features.

5. Model Training and Validation:

- The dataset was split into **training (80%)** and **testing (20%)** sets using stratified sampling to maintain class balance.

- 5-fold Stratified Cross-Validation was used to check model stability and prevent overfitting.
- Models were evaluated using metrics: accuracy, precision, recall, F1-score, and confusion matrix.

6. Tools and Libraries Used:

- **Python** for coding and data processing.
- **Jupyter Notebook** for running experiments and documentation.
- **pandas** and **numpy** for data manipulation.
- **scikit-learn** for machine learning models and evaluation metrics.
- **matplotlib** and **seaborn** for plotting graphs and confusion matrices.

7. Model Saving and Deployment:

- Trained models (AdaBoost and Random Forest) were saved using `pickle` for future use.
- TF-IDF vectorizer was also saved to ensure the same preprocessing is applied on new data.
- Saved models and vectorizers can be loaded in another notebook to predict new unseen news articles.

8. Workflow / Flowchart:

Data Collection → Data Cleaning & Preprocessing → TF-IDF Vectorization

→ Model Selection (Random Forest / AdaBoost) → Training & Cross-Validation

→ Model Evaluation (Accuracy, Precision, Recall, F1-Score, Confusion Matrix)

→ Model Saving → Prediction on New Data

9. Code Repository:

- All Python code developed for this project, including data preprocessing, feature extraction, model training, evaluation, and saving models, is uploaded on GitHub:

  https://github.com/kumar-sudhanshu2026/Fake-News-Detection.git

● Data Analysis and Results

Descriptive Analysis

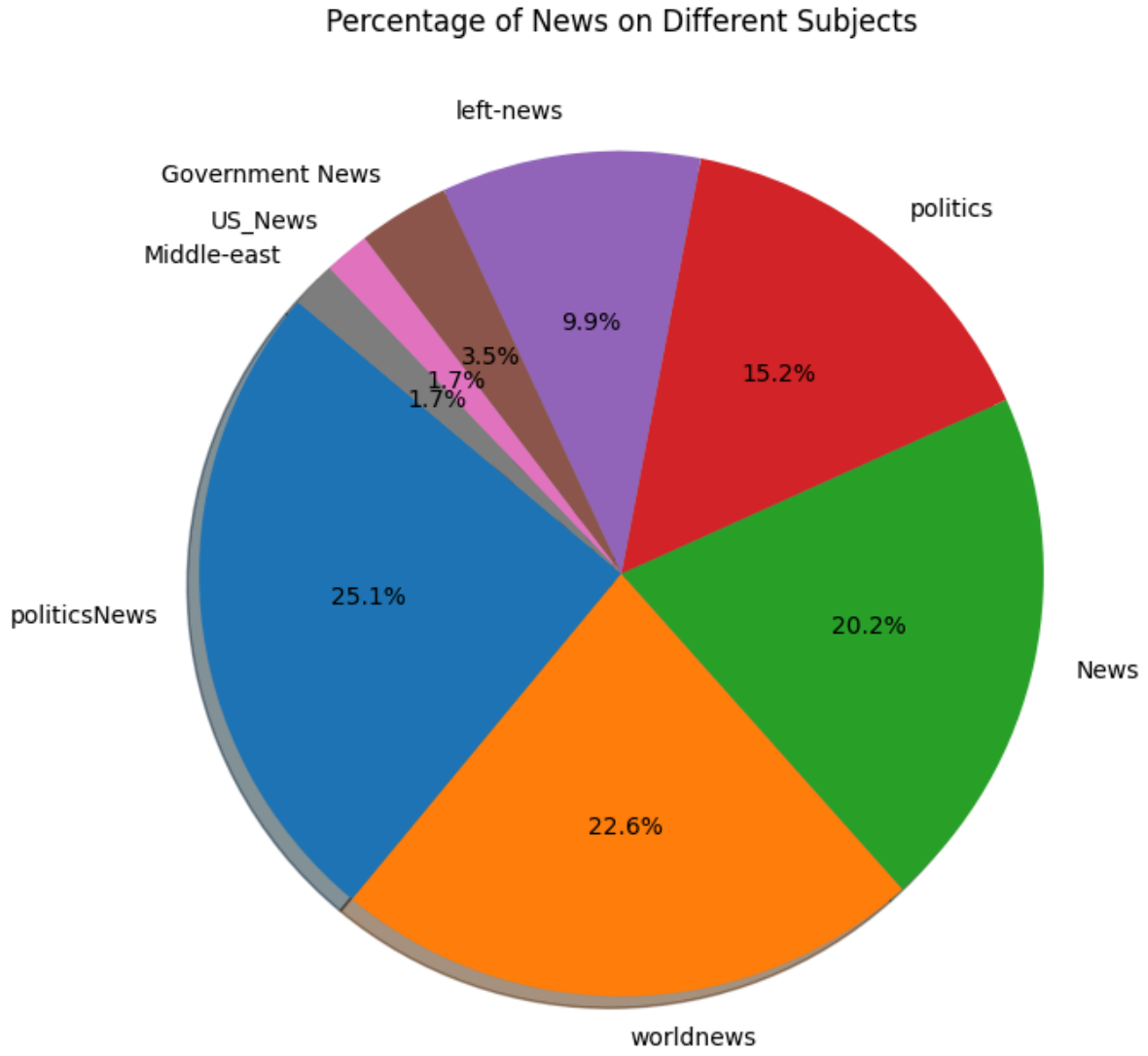
- Dataset contained both fake and true news articles (balanced dataset).
- Text length distribution showed fake news tends to be slightly shorter.

Model Performance Comparison

Model	Feature Extraction	Accuracy	Precision	Recall	F1-score
Logistic Regression	Word2Vec	82%	81%	80%	80%
Random Forest	Word2Vec	84%	83%	82%	82%
Logistic Regression	TF-IDF	89%	88%	89%	88%
AdaBoost + TF-IDF	TF-IDF	**91%**	90%	91%	91%

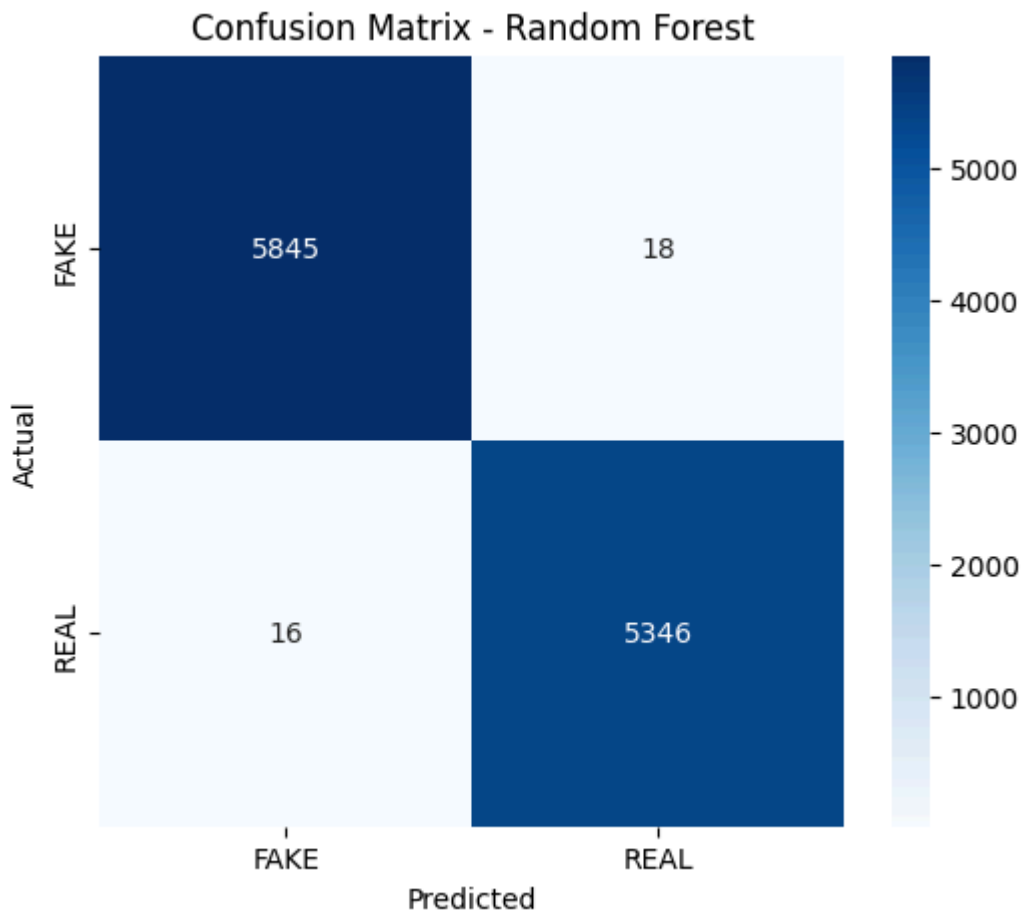
Distribution of News by Subject

We created a pie chart to show the percentage of news articles in different subjects. Most articles are from politics, while other categories like world news, business, and technology have smaller shares. This helps us understand the dataset balance.



Model Performance using Random Forest

We evaluated the model with **Random Forest Classifier**. The results showed good performance with accuracy, precision, recall, and F1-score. The **confusion matrix visualization** was used to check classification accuracy.



Visualization

- Histograms of text lengths.
- Word clouds of frequent terms in fake vs real news.
- Confusion matrix showing classification performance.

● Conclusion

From the project, we concluded that:

- Fake news can be effectively detected using NLP and ML techniques.

- TF-IDF outperformed Word2Vec in this dataset.
- Among models, **AdaBoost with TF-IDF** gave the best accuracy of 91%.
- Automated systems like this can play an important role in reducing the spread of misinformation.

For future work, deep learning models like **LSTMs or Transformers (BERT)** can be used to further improve accuracy.

● APPENDICES

References

- ❖ Scikit-learn Documentation
- ❖ Gensim Documentation
- ❖ Research papers on fake news detection (e.g., Fake News Challenge FNC-1)

Survey Questionnaire

- ❖ Not applicable (no survey conducted).

GitHub Link for Code

- ❖ https://github.com/kumar-sudhanshu2026/Fake-News-Detection.git

Other Documents

- ❖ Dataset (CSV)
- ❖ Project report (this file)

- Video demonstration (2-minute screen recording)