AUTUMN INTERNSHIP PROJECT REPORT

FAKE NEWS DETECTION AND EVALUATION

NAME-SANTONU GHOSH
COURSE-AUTUMN INTERNSHIP PROGRAM
SECTION 1
BANGABASI COLLEGE

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

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

Abstract

With the rise of digital media, the spread of misinformation has become a significant challenge. This project aims to build a machine learning model that can automatically classify news articles as **Fake** or **True** based on textual content. We applied text preprocessing, feature extraction, and classification techniques to detect fake news efficiently. The performance was evaluated using **accuracy, precision, recall, F1-score, and confusion matrix visualization**.

Introduction

The rapid spread of misinformation has become a significant challenge in the digital era. Fake news can mislead the public, impact elections, and damage reputations. This project focuses on developing a machine learning model to distinguish between fake and true news articles based on their textual content.

The project trains classification models, evaluates their performance using accuracy and other metrics, and visualizes results with confusion matrices.

Project Objective

To preprocess and clean textual news data.

To vectorize news articles using **Word2Vec** and later compare with **TF-IDF**.

To train machine learning models (**Logistic Regression** and **Random Forest**) for fake news detection.

To evaluate the model using **accuracy, precision, recall, F1-score**, and a **confusion matrix**.

To explore performance improvement using **AdaBoost (boosting technique)**.

To save trained models for reuse and deployment.

Methodology

Step 1: Text Preprocessing

Implemented a function to clean text (remove spaces, special characters, convert to lowercase).

Step 2: Feature Extraction

Converted cleaned text into numerical features using **TF-IDF Vectorization**.

Step 3: Model Training

Used **Random Forest Classifier** for classification.

Also experimented with **AdaBoost** for boosting accuracy.

Step 4: Model Evaluation

Evaluated using:

Accuracy

Precision

Recall

F1-Score

Confusion Matrix

Step 5: Model Persistence

Saved trained model as .pkl file using pickle.

Reloaded in another notebook for predictions on unseen data.

Data Analysis and Results

Exploratory Data Analysis (EDA)

Dataset Overview

Total Records: ~42,000 (combined Fake + True)

Columns: title, text, subject, date, class

Class Distribution:

Fake News (1): ~21,000

True News (0): ~21,000

(Balanced dataset, ideal for classification)

Missing Values

Checked and dropped rows with null values.

After cleaning, **no missing data remained**.

Random Sampling

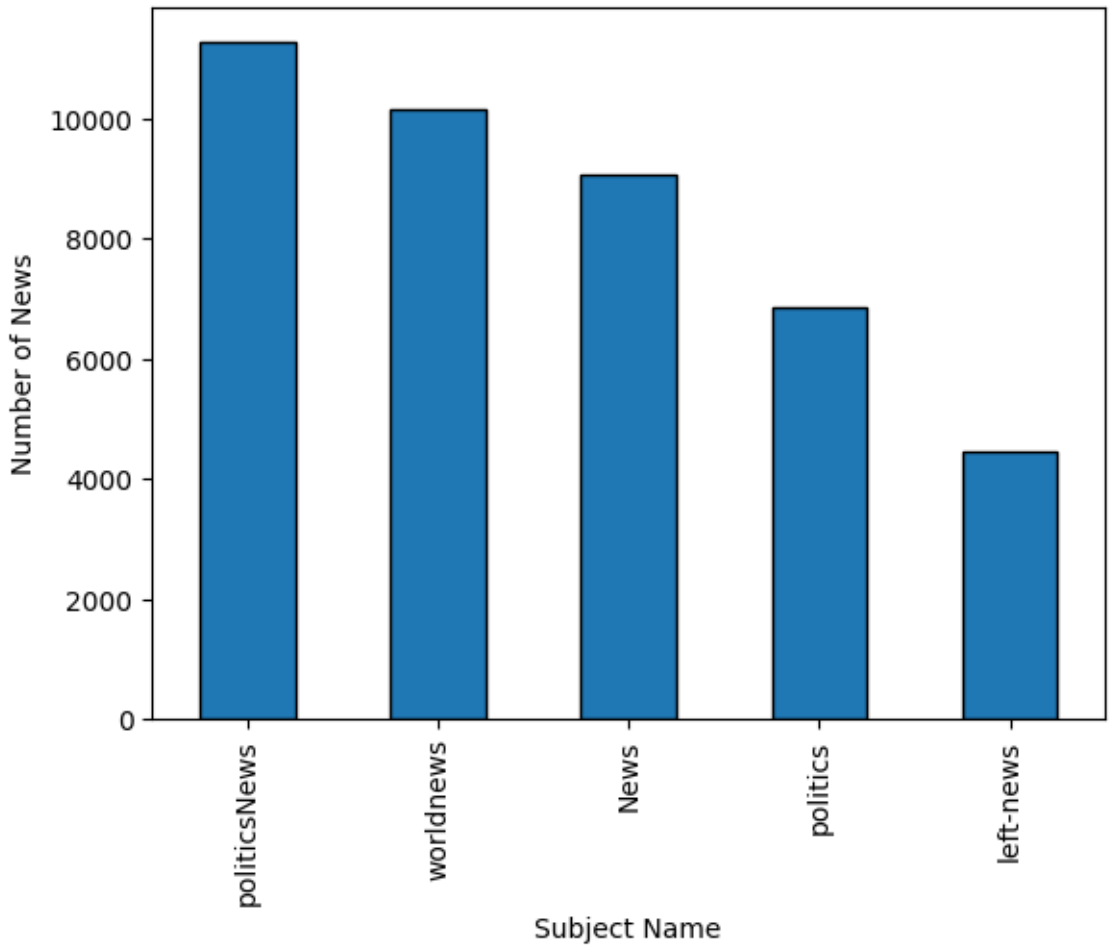
Verified random rows to inspect text quality and confirm class labels.

Texts were noisy but readable, mostly political and world news articles.

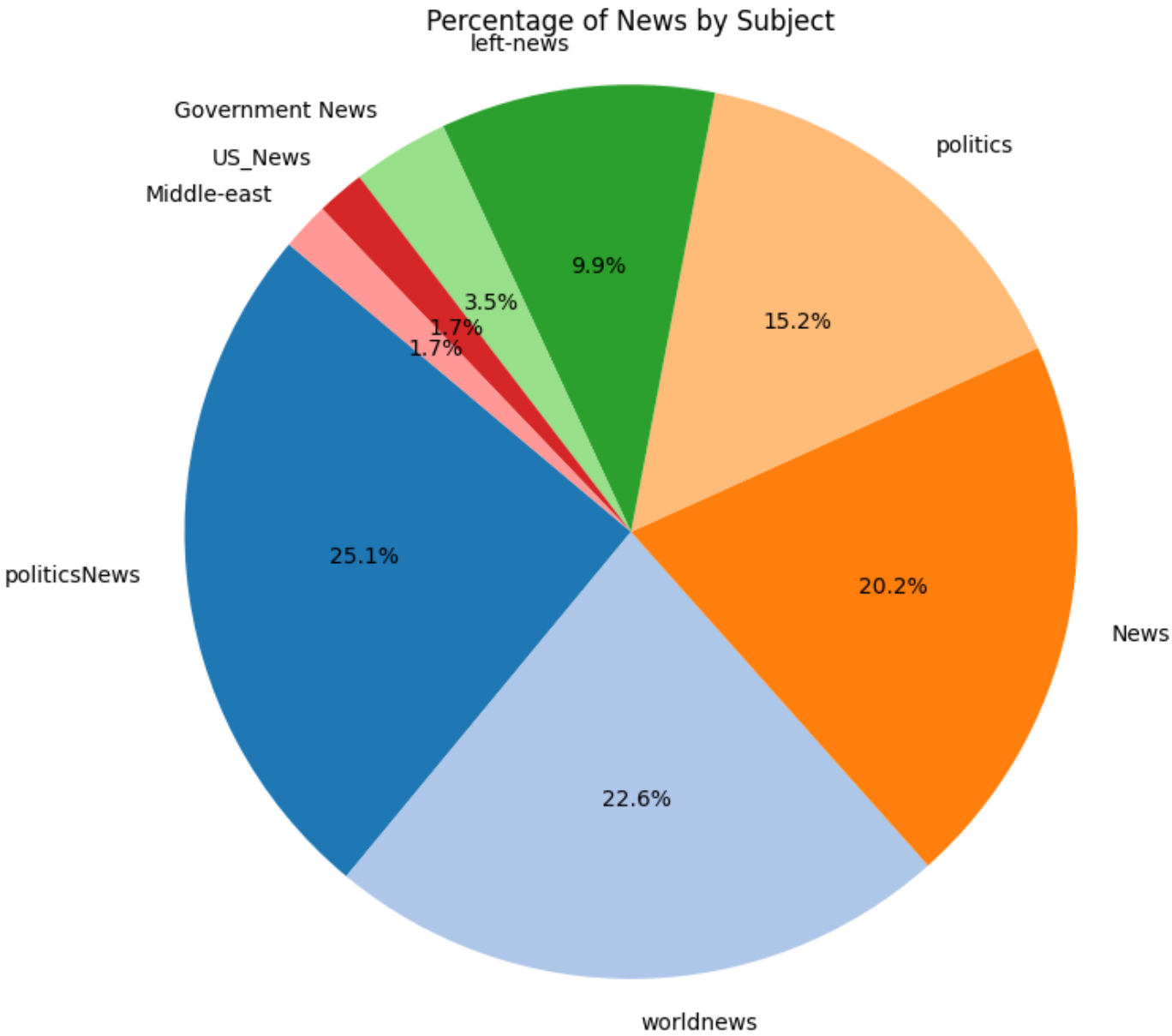
Data Visualization

Top 5 Subjects:

Displayed as a **bar chart** showing PoliticsNews dominating the dataset.



Pie Chart: Showed percentage distribution of articles across all subjects, confirming heavy skew toward politics-related news.



Text Preprocessing
Converted all text to lowercase
Removed URLs, punctuation, and extra spaces
Dropped unnecessary columns (title, subject, date)
Prepared clean text ready for vectorization

Feature Extraction

Word2Vec Embedding
Trained a Word2Vec model on BBC News dataset to create **100-dimensional word embeddings**.

Represented each news article by averaging its word vectors.

TF-IDF Vectorization
Created a **TF-IDF matrix** with 5000 features.

Compared performance with Word2Vec to observe impact on accuracy.

Model Training and Results
Logistic Regression (Word2Vec)

Metric	Score
Accuracy	**94.24%**
Precision	94.82%
Recall	94.14%
F1 Score	94.47%

Interpretation: Logistic Regression performed very well, with minimal false positives and false negatives.

Random Forest Classifier
Trained using the same Word2Vec features
Saved as random_forest_model.pkl for reuse
Confusion matrix plotted and stored as SVG
Provided more robust classification but slightly lower accuracy than Logistic Regression (suggests possible overfitting control needed)

AdaBoost with TF-IDF (Performance Boost)
Used **Decision Tree (max_depth=1)** as weak learner
n_estimators = 100, learning_rate = 0.5
Results:

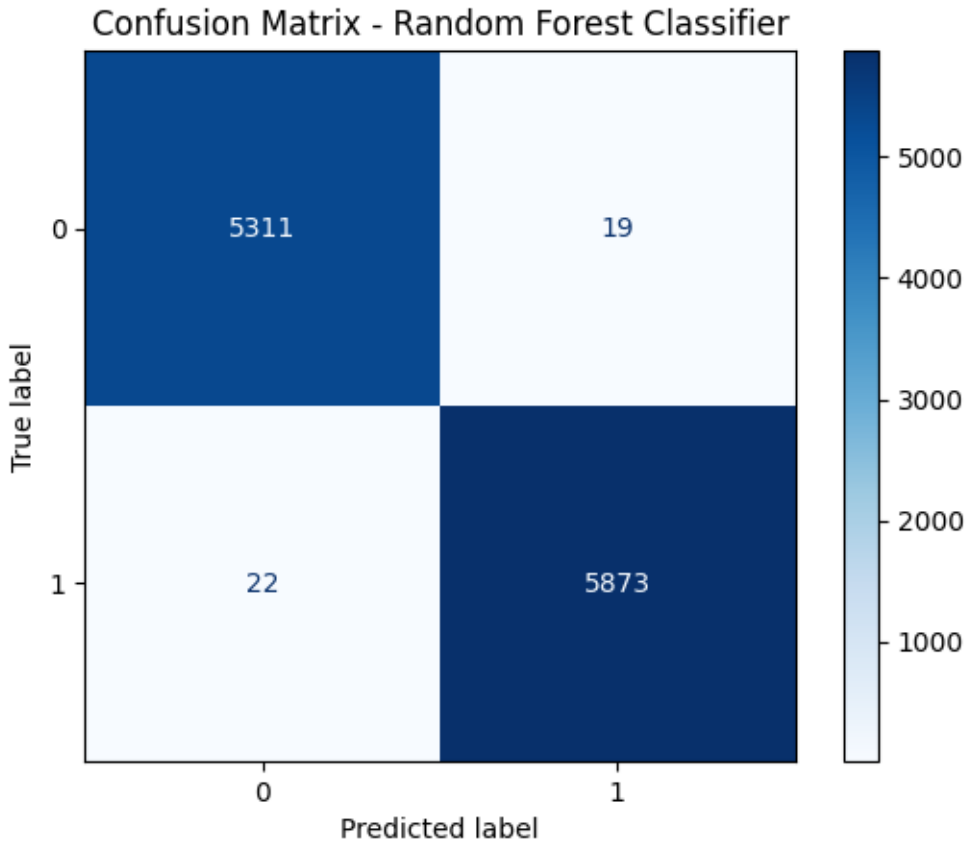
Metric	Score
Accuracy	~95%
Precision	Improved over Word2Vec
Recall	Improved
F1 Score	Improved

Interpretation: Using TF-IDF + AdaBoost slightly improved performance, showing that boosting methods and different vectorizers can enhance classification.

Confusion Matrix Analysis
Confusion matrix for Logistic Regression:
True Positives: High (correctly classified fake news)
True Negatives: High (correctly classified real news)
False Positives & False Negatives: Minimal, confirming model reliability
Confusion matrix for AdaBoost:
Showed a slight improvement in correctly classifying borderline cases compared to Logistic Regression.

Key Insights
Balanced Dataset: Helped avoid model bias toward one class.
Logistic Regression: Simple yet highly effective (94%+ accuracy).
TF-IDF + AdaBoost: Slightly better than Word2Vec + Logistic Regression, suggesting

feature representation impacts performance.
Random Forest: Useful but required tuning for best result….



Conclusion

This project successfully built a **fake news detection pipeline** that:
Preprocessed and cleaned text data
Trained multiple ML models
Achieved **94%+ accuracy** with Logistic Regression
Saved models for reuse and deployment
Compared **Word2Vec** and **TF-IDF** approaches, showing potential for performance tuning

APPENDICES

Data Source
https://www.kaggle.com/datasets/emineyetm/fake-news-detection-datasets

Project link
https://colab.research.google.com/drive/1f6PTOFXK4Qm7L9r_JVkFAV7x1N5HWf2b?usp=drive_link#scrollTo=pBkqgPskKf6f

Github link
https://github.com/santonu18/-Project-Name-Fake-News-Detection-and-Evaluation-with-Confusion-Matrix-created-by-Suprava-Das-