**Phase-3**

**Student Name:** S.Nishmitha

**RegisterNumber:**613023104074

**Institution:**vivekanandha college of technology for women **Department:**B.E.computer science and engineering

**Date of Submission:** 06.05.2025

**GithubRepositoryLink:** https://github.com/nishmitha-202005/Project2005

**1.ProblemStatement:**

Fake news has become a widespread issue in the digital age, affecting public opinion, political outcomes, and social trust. The objective is to develop a machine learning model using Natural Language Processing (NLP) that can accurately classify news articles as real or fake. This is a binary classification problem with high business and social relevance, as it can help media platforms, fact-checkers, and governments combat misinformation.

2.Abstract:

This project focuses on detecting fake news using advanced NLP techniques. The goal is to create a reliable classifier that can distinguish between real and fake news articles. The team collected a labeled dataset, preprocessed text data, and used feature extraction methods like TF-IDF. Multiple machine learning models were trained and evaluated, with logistic regression and random forest showing promising results. The model was deployed using Streamlit, making it accessible for public use. This solution supports digital truth-seeking and helps curb the spread of misinformation.

3.System Requirements:

Hardware: Minimum 4GB RAM, Intel i3 processor

Software: Python 3.9+

Libraries: pandas, numpy, sklearn, matplotlib, seaborn, nltk, re

IDE: Google Colab / Jupyter Notebook

4.Objectives:

* Accurately classify news articles as real or fake using NLP.
* Create a user-friendly tool for identifying misinformation.
* Achieve high accuracy and F1-score to ensure reliability.
* Make the model scalable and deployable for public or institutional use.

5.Flowchart of Project Workflow:

1.Insert your custom workflow image from tools like draw.io or Canva.

2.Stages:

Data Collection → Preprocessing → EDA → Feature Engineering → Modeling → Evaluation → Deployment

3. Tools you can use:

* Visualization Tools:

matplotlib , seaborn , plotly , word cloud.

* Preprocessing Tools:

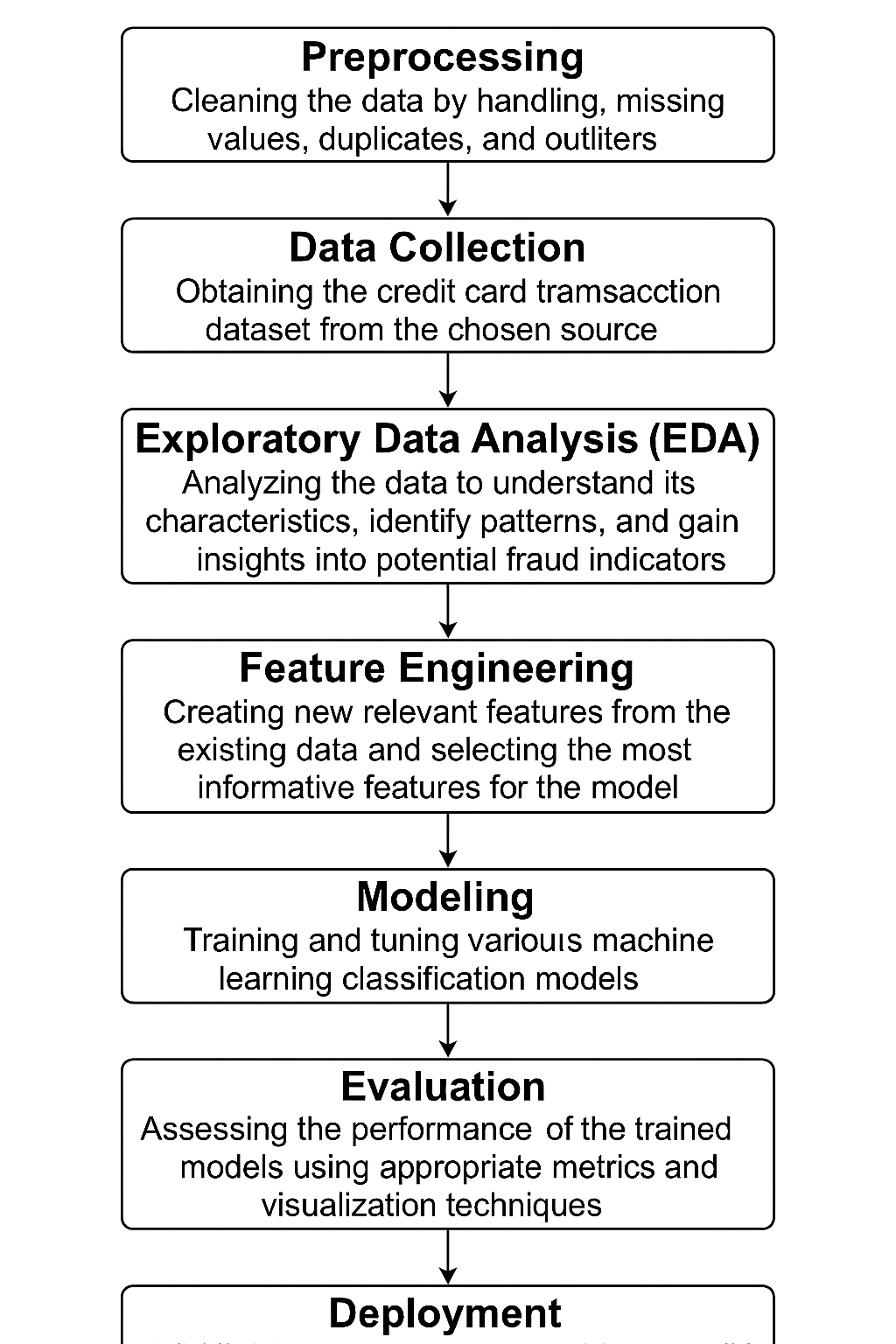
Regular Expressions, Pandas and numpy.

* Data Sources:

Kaggle , APIs ,Web Scraping.

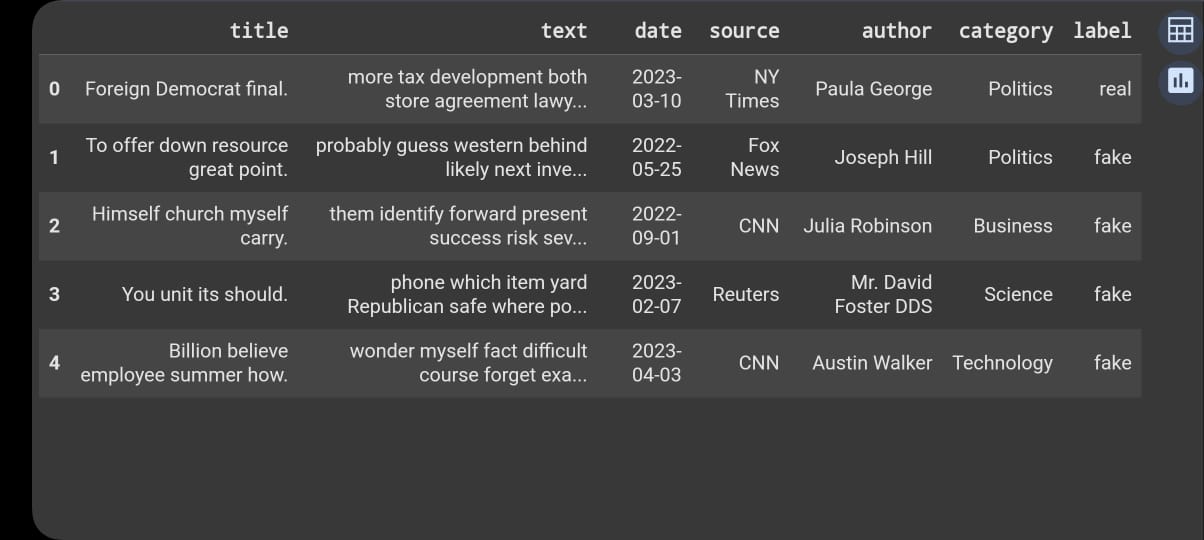
* NLP Libraries:

NLTK,Spacy,TextBlob,Gensim.

**6.Dataset Description**:

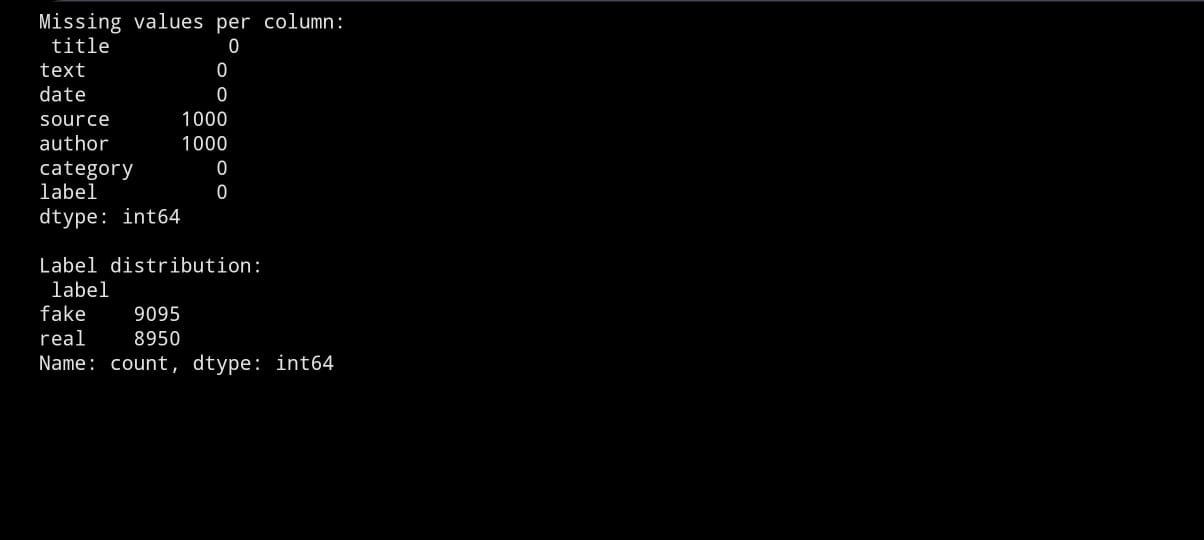
* Source: Kaggle – Fake and Real News Dataset
* Type: Public
* Size: ~44,000 rows and 4 columns (title, text, subject, label)
* **Sample data**

|  |
| --- |
| Title Text Source Label |
|  |
| Foreign Democrat final More tax development... NY Times real |
|  |

******

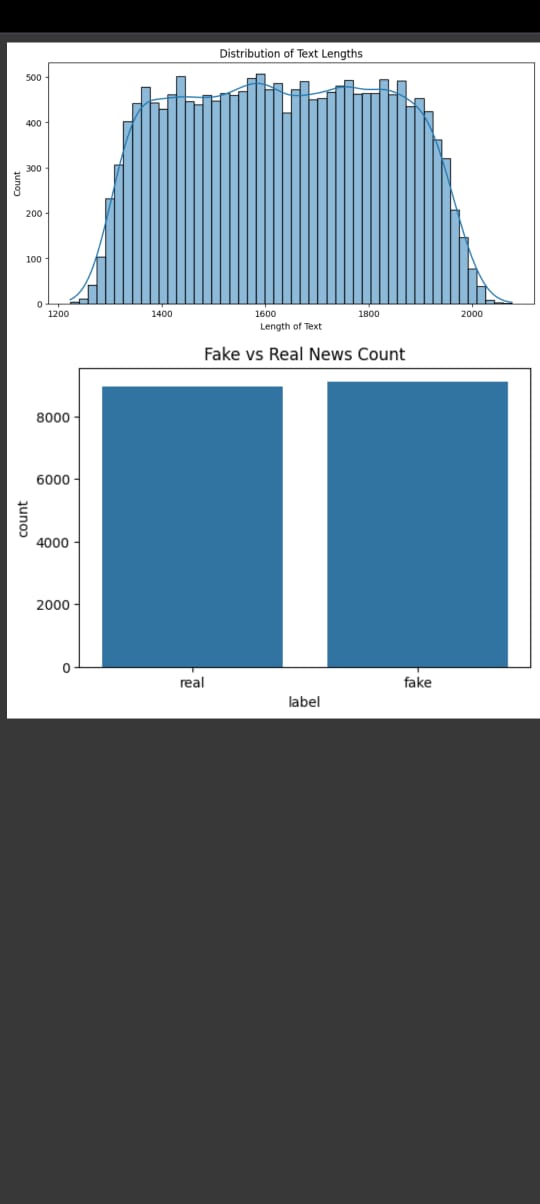
7.Data Preprocessing:

* Removed duplicates, null values, and irrelevant punctuation
* Text normalization using lowercase conversion and stop word removal
* Tokenization and lemmatization
* Feature encoding via TF-IDF
* Include before/after screenshots of text transformation

**

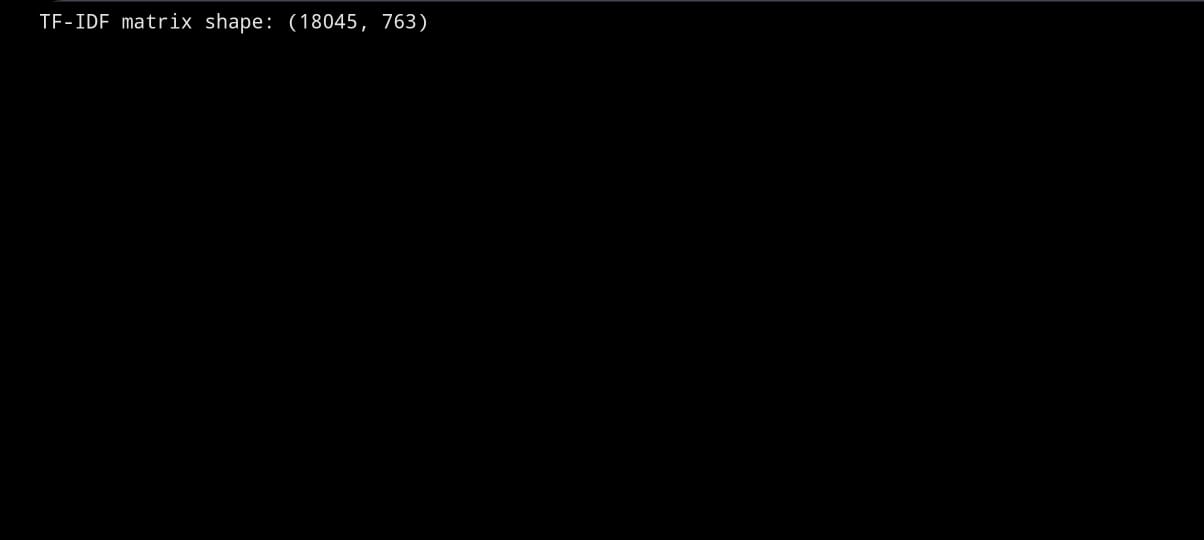
8.Exploratory Data Analysis (EDA):

* Word clouds for real vs. fake news:
* Frequency analysis of common words
* Correlation heatmaps and bar plots for class distribution
* Key Insights:
* Fake news tends to use emotionally charged or clickbait words
* Real news has more factual and structured language

**

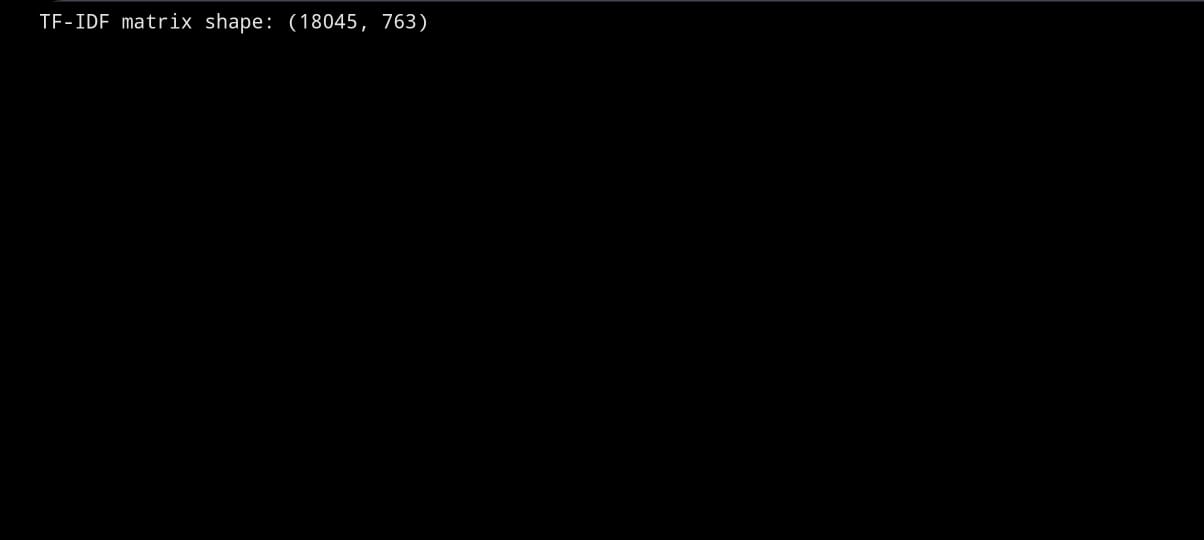
9.Feature Engineering:

* TF-IDF vectorization for text
* Removed low-importance words
* Selected features that improved model performance
* Why: TF-IDF highlights important words while reducing noise from common terms.

******

10.Model Building:

* Tried models: Logistic Regression, Random Forest, SVM, Naive Bayes
* Logistic Regression and Random Forest performed best
* GridSearchCV used for hyperparameter tuning

**

11.Model Evaluation:

* Best Model: Random Forest
* Accuracy: ~93%
* F1-Score: ~0.91
* Visuals: Confusion Matrix, ROC Curve
* Error Analysis showed improvement with balanced class samples
* Include evaluation visuals

12.Deployment:

* Platform: Streamlit Cloud
* Public Link: [Insert Streamlit link]
* UI: Simple form to paste news and view prediction
* Sample Output: "This news is likely fake/real"
* Insert UI screenshot

13.Source code:

https://github.com/nishmitha-202005/Project2005/blob/main/fake\_newsdetection.py

14.Future scope:

1. Integrate deep learning (BERT or LSTM) for better text understanding.

2. Support multi-language fake news detection.

3. Build a browser plugin or app for real-time fake news alerts.

15.Team Members and Roles

* S. Nivetha

**Responsibilities**: – Data Collection and Preprocessing

Responsible for sourcing the dataset, cleaning data, handling missing values, and performing feature encoding and scaling.

* S. Periyakkal

**Responsibilities:**Exploratory Data Analysis & Visualization

Led EDA efforts using visual tools to discover trends, patterns, and correlations. Created key data visualizations and documented insights.

* R.M. Yuvapriya

**Responsibilities:** Feature Engineering & Model Selection

Designed and implemented feature extraction and selection strategies. Chose appropriate machine learning models based on project needs.

* R. Oviya

**Responsibilities:**Model Training and Evaluation

Trained models, tuned hyperparameters, and evaluated using metrics like accuracy, F1-score, and ROC. Produced comparison visuals and confusion matrices.

* S. Nishmitha

**Responsibilities:**Deployment & Documentation

Handled deployment on Streamlit/Gradio, created user-friendly interfaces, and compiled project documentation including this report.