Phase-2 Submission

Student Name: Priyadharshan P

Register Number: 712523205044

Institution: PPG Institute of Technology

Department: Information Technology

Date of Submission: 02-05-2025

Github Repository Link:
https://github.com/dharshu0623/NM_Priyadharshan_DS

---

1. Problem Statement

In today's digital age, misinformation spreads rapidly across social media platforms, often leading to social, political, and economic consequences. Detecting fake news has become a critical necessity. This project addresses the ***classification problem*** *of identifying whether a news article is real or fake using advanced* ***Natural Language Processing (NLP)*** *techniques.*

Solving this problem is significant because:

- *It helps curb the spread of misinformation.*
- *It enables platforms and users to flag or filter deceptive content.*
- *It aids in media literacy and public awareness.*

2. Project Objectives

- *Develop an NLP-based pipeline to classify news articles as real or fake.*
- *Preprocess and clean unstructured text data to ensure quality input for models.*

- *Apply multiple machine learning algorithms and compare their performance.*
- *Optimize for accuracy, F1-score, and generalizability.*
- *Deliver interpretable results using visualization techniques.*

*After data exploration, we refined our goal to focus more on **explainability** and **robustness** of predictions.*

3. Flowchart of the Project Workflow

4. Data Description

- ***Dataset name**: Kaggle – Fake and Real News Dataset*
- ***Source**: Kaggle – Fake and Real News Dataset*
- ***Type**: Unstructured (Text)*
- ***Records**: ~44,919 articles (23,502 fake, 21,417real)*

- ***Features**: title, text, subject, date*
- ***Target Variable**: label (FAKE = 1, REAL = 0)*
- ***Dataset Nature**: Static*
- ***Dataset Link**: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset?resource=download*

5. Data Preprocessing

- *Removed null or irrelevant entries (e.g., blank titles or text).*
- *Eliminated duplicates and standardized date formats.*
- *Cleaned text:*
    - *Lowercased all text*
    - *Removed punctuation, stopwords, and non-alphanumeric tokens*
    - *Applied lemmatization (using spaCy)*
- *Encoded label as binary.*
- *Vectorized text using:*
    - *TF-IDF (for baseline models)*
    - *Word embeddings (optional for deep models)*

6. Exploratory Data Analysis (EDA)

***Univariate Analysis**:*

- *Word clouds for fake vs. real news.*
- *Most frequent terms in both classes.*
- *Length distribution of articles.*

***Bivariate Analysis**:*

- *Bar plots comparing subject categories across labels.*
- *Correlation between article length and likelihood of being fake.*

***Insights**:*

- *Fake news tends to use sensational terms ("shocking", "unbelievable").*
- *Titles of fake articles are typically shorter but more exaggerated*

7. Feature Engineering

- *Extracted article length, title length as new numerical features.*
- *Used TF-IDF with unigrams and bigrams.*
- *Created binary features for presence of sensational words.*
- *Optional: Used Latent Semantic Analysis (LSA) for dimensionality reduction.*

8. Model Building

***Models Used**:*

- *Logistic Regression (Baseline)*
- *Random Forest Classifier*
- *Support Vector Machine (SVM)*
- *(Optional) LSTM/Transformer-based model for advanced NLP*

***Why These Models**:*

- *Logistic Regression: Interpretability and efficiency.*
- *Random Forest: Handles noise and non-linear patterns.*
- *SVM: Effective in high-dimensional text spaces.*

***Evaluation Metrics**:*

- *Accuracy*
- *Precision*
- *Recall*
- *F1-Score*
- *ROC-AUC*

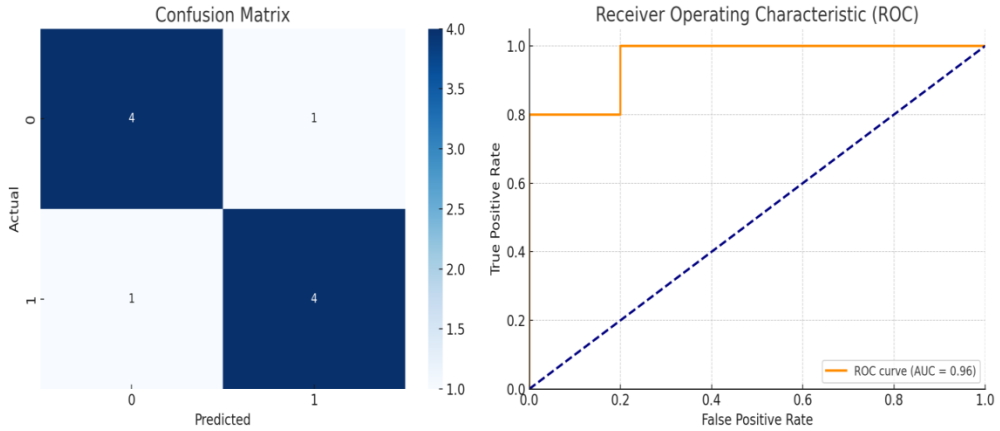
Train-test split: 80:20 with stratification

9. Visualization of Results & Model Insights

- ***Confusion Matrix**: To visualize misclassifications.*
- ***ROC Curve**: AUC scores for model comparison.*
- ***Feature Importance**: For tree-based models.*
- ***Misclassified Samples**: Review of top wrongly predicted articles.*

***Insights**:*

- *TF-IDF + SVM performed best with ~93% accuracy.*
- *Most important features were emotionally charged keywords and certain subject categories.*



10. Tools and Technologies Used

- ***Language**: Python*
- ***IDE**: Jupyter Notebook / Google Colab*
- ***Libraries**:*
  - *Data Handling: pandas, numpy*
  - *Visualization: matplotlib, seaborn, plotly*
  - *NLP: nltk, spaCy, scikit-learn, wordcloud*
  - *ML Models: scikit-learn, xgboost, keras (optional)*
- ***Version Control**: GitHub*

11. Team Members and Contributions

S.no	Name	Role
1	Priyadharshan P	Data cleaning
2	Bindhiya T.	Model development
3	Akhilan B.	Documentation and reporting
4	James Aathithyan A.	EDA
5	Anish M.	Feature engineering