Phase-2 Submission

Student Name:Bharathidhasan M

Register Number: 712523104007

Institution: PPG INSTITUTE OF TECHNOLGY

Department: BE CSE

Date of Submission: 08-05-2025

Github Repository:
https://github.com/bharathidhasan26/NM_bharathidhasan__DS.git

---

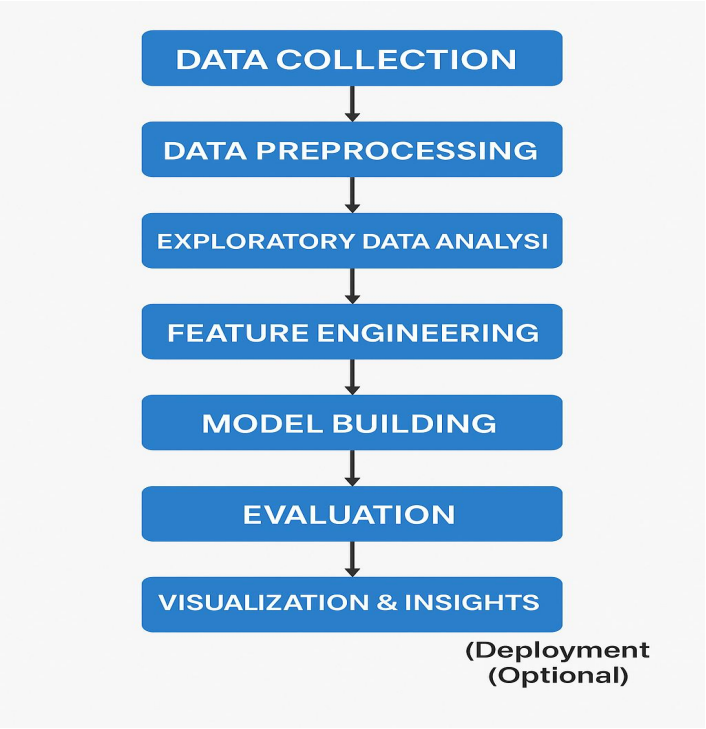
1. Problem Statement

The proliferation of fake news on social media and news websites poses a significant threat to societal trust, public safety, and democracy.It is framed as a ***binary classification*** *problem, where the goal is to label a news article as either* ***real*** *or* ***fake*** *based on its textual content. Solving this problem enables platforms to flag potentially misleading content, making online information more reliable.*

2. Project Objectives

- *Build a supervised machine learning model to classify news as real or fake.*
- *Use NLP techniques to process and extract features from raw text data.*
- *Evaluate and compare multiple classification models for performance.*
- *Visualize insights into model performance and feature relevance.*

3.Flowchart of the Project Workflow;



4. Data Description ;

- *Dataset Source: Kaggle - "Fake and Real News Dataset"*
- *Type of Data: Unstructured text (news articles)*
- *Records & Features: ~44,000 articles with columns like title, text, label*
- *Static Dataset: Yes*
- *Target Variable: label (1 for Real, 0 for Fake)*

DATA SET : https://www.kaggle.com/datasets/mahdimashayekhi/fake-new

	Title	Text	Date	Source	Author	Category	Label
2	Foreign Democrat	more tax (	10-03-2023	NY Times	Paula Geo	Politics	real
3	To offer down res	probably (	25-05-2022	Fox News	Joseph Hil	Politics	fake
4	Himself church my	them iden	01-09-2022	CNN	Julia Robir	Business	fake
5	You unit its shoulc	phone wh	07-02-2023	Reuters	Mr. David	Science	fake
6	Billion believe emp	wonder m	03-04-2023	CNN	Austin Wa	Technolog	fake

5. Data Preprocessing;

- *Merged title and text to form a complete input for analysis.*
- *Removed duplicate articles.*
- *Cleaned text: lowercased, removed punctuation, stopwords, and digits.*
- *Applied lemmatization to reduce word forms.*
- *Encoded target variable (label) using binary encoding*

6. Exploratory Data Analysis (EDA)

Univariate Analysis:

- *Word clouds of most common words in fake vs. real articles.*
- *Distribution of article lengths.*

Bivariate/Multivariate Analysis:

- *Countplot of label distribution revealed mild class imbalance.*
- *N-gram frequency analysis for fake and real news.*

Insights Summary:

- *Fake news often uses more sensational and emotionally charged language.*
- *Real news tends to include names of known organizations or references to factual events.*
- *Features like word count and specific keywords were found to influence*

classification.

7. Feature Engineering

- *Combined title and text into one feature: full_text.*
- *Extracted article length and average word length as new numerical features.*
- *TF-IDF applied to extract key textual patterns.*
- *Removed very frequent and very rare words (min_df=5, max_df=0.8) to reduce noise.*

8. Model Building

- ***Logistic Regression*** *– baseline model for interpretability.*
- ***Random Forest Classifier*** *– for handling high-dimensional data and nonlinear patterns.*
- ***Multinomial Naive Bayes*** *– commonly used for text classification.*
- ***4. Model Evaluation Metrics:****Accuracy, Precision, Recall, F1-Score*
- *Cross-validation used for better generalization.*
- *Random Forest performed best with an F1-score of ~0.93 on test data.*

9. Visualization of Results & Model Insights

- ***Confusion Matrix:*** *Showed true positives, false positives, etc.*
- ***ROC Curve:*** *AUC of 0.96 for the best model.*

- ***Feature Importance (Random Forest):*** *Showed top TF-IDF words contributing to classification.*
- ***Word Clouds:*** *Differentiated language styles of fake vs. real news.*

10. Tools and Technologies Used

- *Programming Language: Python*
- *IDE/Notebook: Google Colab*
- *Libraries: pandas, numpy, scikit-learn, nltk, seaborn, matplotlib, xgboost, plotly*
- *Visualization Tools: matplotlib, seaborn, wordcloud, plotly*

11. Team Members and Contributions

NAME	RO LE	WORK
MURALIDHARAN K	Team Coordinator	
GOWTHAM P	Marketing & Outreach Lead	
PUGAZHENTHI	NLP Engineer	
BHARATHIDHASAN	Document and presentation	
JESLIN SAJAN	Testing and deployment	