**Phase-1 Submission**

**Student Name:** Priyadharshan P

**Register Number:** 712523205044

**Institution:** PPG Institute of Technology

**Department:** Information Technology

**Date of Submission:** 25-04-2025

**1. Problem Statement**

*With the widespread use of social media and online platforms, fake news has become a major threat to societal trust and informed decision-making. People are easily misled by false narratives, which can cause misinformation to spread rapidly. This project aims to develop a robust fake news detection system that leverages Natural Language Processing (NLP) to identify and filter deceptive content from reliable information sources.*

**2. Objectives of the Project**

* *To detect fake news articles based on their textual content using advanced NLP techniques.*
* *To develop a machine learning model that classifies news as "Fake" or "Real" with high accuracy.*
* *To analyze and understand the linguistic patterns and markers that distinguish real news from fake.*
* *To deliver insights through visualizations and deploy a user-friendly interface for fake news verification*

**3. Scope of the Project**

***Inclusions:***

* *Use of NLP for text preprocessing and feature extraction.*
* *Implementation of machine learning models (e.g., Logistic Regression, SVM, Random Forest, etc.).*
* *Comparative evaluation of model performance.*
* *Visualization of dataset insights and model predictions.*

***Limitations:***

* *Focused only on English language text.*
* *Static dataset usage (not real-time crawling).*
* *Not covering multimedia (image or video-based fake news).*

**4. Data Sources**

* ***Dataset:*** *"Fake and Real News Dataset" from Kaggle.*
* <https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification>
* <https://www.kaggle.com/datasets/bhavikjikadara/fake-news-detection>
* <https://www.kaggle.com/datasets/vishakhdapat/fake-news-detection>
* <https://zenodo.org/records/6555293>
* [https://huggingface.co/datasets/ErfanMoosaviMonazzah/fake-news-detection-dataset-English](https://huggingface.co/datasets/ErfanMoosaviMonazzah/fake-news-detection-%20%20%20%20%20%20%20%20%20%20%20dataset-English)
* <https://data.niaid.nih.gov/resources?id=zenodo_11408512>
* <https://data.niaid.nih.gov/resources?id=zenodo_10354244>
* ***Nature:*** *Static dataset downloaded once.*
* ***Details:*** *Contains labeled news articles with metadata such as title, text, and source.*

**5. High-Level Methodology**

* **Data Collection** – *Download the dataset from Kaggle.*
* **Data Cleaning**– *Remove stop words, punctuation, and HTML tags & Handle missing values and duplicates*
* **Exploratory Data Analysis (EDA)** – *Use word clouds, bar plots, and correlation matrices to uncover trends & Analyze word frequency for fake vs. real news.*
* **Feature Engineering** *– Use TF-IDF and Count Vectorizer for feature extraction & Consider n-grams and word embeddings for better representation*.
* **Model Building** – *Apply and compare models: Logistic Regression, Naive Bayes, SVM and Random Forest. Use cross-validation to prevent over fitting.*
* **Model Evaluation** – *Metrics: Accuracy, Precision, Recall, F1-score, Confusion Matrix & ROC-AUC Curve for model comparison.*
* **Visualization & Interpretation** – *Charts to show accuracy comparison, feature importance, and EDA insights & Interpret model decisions using SHAP or LIME if applicable*.
* **Deployment** – *Deploy using Streamlit for a simple web-based news classifier. Optional integration of a text input box for real-time prediction.*

**6. Tools and Technologies**

* **Programming Language** – *Python*
* **Notebook/IDE** – *Google Colab / Jupyter Notebook*
* **Libraries** – *Data Processing: pandas, numpy*

*NLP: nltk, spacy, scikit-learn*

*Visualization: matplotlib, seaborn, wordcloud*

*Modeling: scikit-learn*

* **Optional Tools for Deployment** – *Streamlit, Flask*

**7. Team Members and Roles**

|  |  |  |
| --- | --- | --- |
| ***S.no*** | ***Name*** | ***Role*** |
| *1* | *Priyadharshan P* | *Project Lead, Model Development* |
| *2* | *Bindhiya T.* | *Data Collection & EDA* |
| *3* | *Akhilan B.* | *Visualization & Presentation* |
| *4* | *James Aathithyan A.* | *Deployment & UI Development* |
| *5* | *Anish M.* | *Feature Engineering & Model Evaluation* |