**Phase-2 Submission Template**

**Student Name:** M.Saravanan

**Register Number:** 621523104143

**Institution:** Mahendra College of Engineering

**Department:** Computer Science and Engineering

**Date of Submission:** 08/05/2025

**GithubRepositoryLink:** thehttps://github.com/Selvabharathy13



### **1. Problem Statement** *With the exponential growth of online content, fake news has become a significant threat to public opinion, political stability, and public health. Misinformation spreads faster than ever, misleading readers and creating confusion. There is an urgent need for intelligent systems that can accurately detect and flag fake news using computational approaches like Natural Language Processing (NLP) and Machine Learning.*

### **2. Project Objectives**

* *1. To collect and preprocess a dataset consisting of labeled real and fake news articles.*
* *2. To analyze the linguistic and statistical patterns that differentiate fake news from real.*
* *3. To engineer relevant textual features using NLP techniques.*
* *4. To build and compare multiple machine learning models for fake news classification*

### **3. Flowchart of the Project Workflow**

### *┌────────────────────┐*

### *│ Data Collection │*

### *└────────┬───────────┘*

### *↓*

### *┌────────────────────┐*

### *│ Data Preprocessing│*

### *└────────┬───────────┘*

### *↓*

### *┌───────────────▼────────────────┐*

### *│ Exploratory Data Analysis (EDA)│*

### *└────────┬───────────────────────┘*

### *↓*

### *┌────────▼────────┐*

### *│Feature Engineering│*

### *└────────┬────────┘*

### *↓*

### *┌──────▼──────┐*

### *│Model Building│*

### *└──────┬──────┘*

### *↓*

### *┌───────▼────────┐*

### *│Model Evaluation│*

### *└───────┬────────┘*

### *↓*

### *┌───────▼────────────┐*

### *│ Visualization & Insights│*

### *└────────────────────┘*

### 

### **4. Data Description**

*Dataset Name: Fake and real news dataset*

*Source: Kaggle or other news corpus datasets*

*Attributes:*

*title: Headline of the article*

*text: Full article content*

*subject: Topic/category of the news*

*date: Publication date*

*label: 0 for real, 1 for fake*

### **5. Data Preprocessing**

### Removing null values

### Dropping duplicates

### Cleaning text (lowercasing, removing punctuation, stopwords, and numbers)

### Tokenization and lemmatization

### Balancing the dataset (if imbalanced)

### **6. Exploratory Data Analysis (EDA)**

*Word frequency analysis*

*Word clouds for fake vs real news*

*Article length distribution*

*Frequency of top keywords*

*Distribution of topics (subject)*

*Temporal patterns if using date*

### 

### **7. Feature Engineering**

### Bag of Words (BoW)

### TF-IDF Vectorization

### N-grams (bigrams/trigrams)

### Sentiment scores

### POS tag distribution

### Named Entity Recognition (NER) counts

### **8. Model Building**

### *Algorithms used:*

### *Logistic Regression*

### *Naïve Bayes*

### *Support Vector Machine (SVM)*

### *Random Forest*

### *XGBoost*

### *Deep Learning (LSTM or BERT for advanced model)*

### *Performance metrics:*

### *Accuracy*

### *Precision*

### *Recall*

### *F1-Score*

### *ROC-AUC Curve*

### **9. Visualization of Results&Model Insights**

* *Confusion Matrix*
* *ROC Curve*
* *Precision-Recall curve*
* *Feature importance plot (e.g., top TF-IDF features)*
* *Misclassified sample analysis*
* *Comparative bar chart of model performance*

### 

### **10. Tools and Technologies Used**

* *Languages: Python*
* *Libraries: pandas, numpy, scikit-learn, nltk, seaborn, matplotlib, TensorFlow/Keras, XGBoost*
* *NLP Tools: NLTK, spaCy*
* *Development Environment: Jupyter Notebook, Google Colab*
* *Version Control: Github*

### 

### **11. Team Members and Contributions**

1. Selvabharathy K

Led the data preprocessing and feature engineering phase.

Implemented NLP techniques including tokenization, stopword removal, and TF-IDF vectorization.

Collaborated on model building and code optimization.

2. Samuel Paul V

Focused on exploratory data analysis (EDA) and visualization of insights.

Created word clouds, frequency plots, and distribution graphs for both fake and real news.

Assisted in preparing the final project documentation.

3. Saravanan M

Worked on the machine learning model development using logistic regression, SVM, and ensemble methods.

Tuned hyperparameters and evaluated model performance using classification metrics.

Contributed to comparative analysis of models.

4. Sanjeevi V

Took charge of the deep learning implementation using LSTM and BERT.

Managed the integration of NLP pipelines with neural network models.

Helped with testing and deployment strategies.

5. Sakthivel E

Coordinated the project workflow and task management.

Created the flowchart and system design documentation.

Consolidated team contributions and ensured timely progress and submissions.