**Phase-2 Submission Template**

**Student Name:** *GOKULAPRIYA R*

**Register Number:** 212923104029

**Institution:** *ST.JOSEPH COLLEGE OF ENGINEERING*

**Department:** COMPUTER SCIENCE AND ENGINEERING

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

**Github Repository Link:** [**https://github.com/sekar336/project.git**](https://github.com/sekar336/project.git)

### **PROJECT TITLE:**

### **EXPOSING THE TRUTH WITH ADVANCED FAKE NEWS DETECTION POWERED BY NATURAL LANGUAGE PROCESSING**

### 

### **1. Problem Statement**

### *The surge of fake news on digital platforms poses a serious threat to public opinion, trust, and democratic processes. This project aims to combat misinformation by developing an intelligent NLP-driven system capable of detecting fake news in real-time. Leveraging advanced machine learning and natural language processing techniques, the system classifies content as real or fake, thereby assisting users, organizations, and fact-checkers in making informed decisions.*

### 

### **2. Project Objectives**

### *Build an NLP-based system to detect fake news from textual content.*

### *Implement deep learning and transformer-based models for classification.*

### *Integrate real-time detection into web applications for public use.*

### *Evaluate performance using metrics such as precision, recall, F1-score, and accuracy.*

### *Improve model accuracy through continuous feedback and model updates.*

### 

### **3. Project Flowchart**

### 

### 

### **4. Data Description**

### *Source : Public datasets (e.g., Kaggle), Twitter Sentiment & News Content Analysis*

### *Format : CSV*

### *Fields : Article ID, Timestamp, Headline, Content, Source, Label (Real/Fake)*

### *Volume : ~5000 records*

### *Type : Textual news articles and social media posts*

### 

### **5. Data Preprocessing**

### *Cleaning : Removed HTML tags, emojis, special characters*

### *Tokenization & Lemmatization : Used SpaCy*

### *Stopword Removal : Applied to reduce noise*

### *Balancing : Applied SMOTE for class imbalance*

### 

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

### *Label Distribution : Bar plots showing real vs. fake distribution*

### *Common Keywords : Word clouds for real and fake news*

### *Source Reliability : Frequency analysis of fake news by source*

### 

### **7. Feature Engineering**

### *Extracted Features : Content length, punctuation use, sentiment score*

### *Vectorization : TF-IDF and word embeddings (e.g., GloVe)*

### *Selection : Based on mutual information and feature importance*

### 

### **8. Model Building**

### *Fake News Detection Models :*

### *Logistic Regression*

### 

### *Random Forest*

### 

### *Fine-tuned BERT (Transformers)*

### 

### *Training/Testing Split : 80/20*

### 

### *Performance Metrics :*

### *Logistic Regression: Accuracy = 81%, F1-score = 0.78*

### 

### *BERT Fine-tuned: Accuracy = 92%, F1-score = 0.90*

### 

### **9. Results & Visualizations**

### *Confusion Matrix : Detailed breakdown of model predictions*

### *F1-Score Comparison : Across all models*

### *Attention Visualization (for BERT) :*

### *Highlighting influential words in classification*

### *Insights :*

### *BERT significantly outperformed traditional models, especially on nuanced language.*

### 

### *High-precision keywords and linguistic patterns help distinguish fake from real.*

### 

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

### *Languages : Python*

### *IDE/Notebooks : Jupyter, Google Colab*

### *Libraries : pandas, numpy, scikit-learn, matplotlib, seaborn, SpaCy, HuggingFace Transformers*

### *Deployment : Flask, Streamlit*

### 

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

### *Akshaya A : Data Collection, Cleaning, and Preprocessing*

### *Athmika G : Model Development and Optimization*

### *Gokulapriya R : Visualization and Web Deployment*

### 

### 

### 