**Phase-2 SubmissionTemplate**

**Student Name:** [Mubarak Basha S]

**Register Number:** [510623104069]

**Institution:** [C Abdul Hakeem College of Engg & Tech]

**Department:** [Computer science and engineering]

**Date of Submission:** [09/05/2025]

**Github Repository Link**

https://github.com/MubarakBashaS/Fake-news-detection

# Problem Statement:

In an era dominated by rapid information exchange, the spread of fake news on digital platforms poses a significant threat to public trust, political stability, and societal harmony. The objective of this project is to build an automated system to detect fake news using Natural Language Processing (NLP). This is a binary classification problem where the model predicts whether a given news article is 'real' or 'fake'.

Fake news can mislead the public and incite misinformation at scale. By leveraging advanced NLP techniques, we aim to develop a robust solution that can be applied in journalism, social media moderation, and public safety initiatives.

# Project Objectives

*To preprocess and clean a text-based fake news dataset.*

*To conduct comprehensive EDA to understand linguistic patterns in fake vs real news.*

*To build and evaluate machine learning models for text classification.*

*To optimize model accuracy, precision, and recall while maintaining interpretability.*

To explore the use of advanced NLP models (like TF-IDF, Word2Vec, BERT) for performance comparison.

# Flowchart of the Project Workflow

Data Collection → Data Preprocessing → EDA → Feature Engineering → Model Building → Evaluation → Visualization of Results → Conclusion

# Data Description

*Source: Kaggle – “Fake News Detection” dataset*

*Type: Text (unstructured)*

*Records: ~20,000 news articles*

*Features: Title, Text, Subject, Date, Label*

*Target Variable: Label (1 = Fake, 0 = Real)*

*Nature: Static dataset*

# Data Preprocessing

Removed null values and rows with missing titles or text.

Removed duplicates based on identical article text.

Tokenized and lemmatized text content.

Removed stop words, punctuation, and irrelevant characters.

Converted text to lowercase.

Encoded target variable using label encoding.

# Exploratory Data Analysis (EDA)

*Univariate Analysis:*

*Word clouds for fake and real articles showed stark vocabulary differences.*

*Length of text and title were shorter in fake news.*

*Bivariate Analysis:*

*Real news showed more formal vocabulary.*

*TF-IDF score distribution showed distinguishable patterns across classes.*

*Insights:*

*Fake news tends to use exaggerated and sensational words.*

*Real news contains more references and neutral tone.*

# Feature Engineering

*Applied TF-IDF vectorization for extracting features from text.*

*Experimented with Word2Vec embeddings.*

*Combined title and text fields to improve context understanding.*

Created feature indicating the presence of sensational keywords.

# Model Building

*Models used:*

*Logistic Regression (baseline)*

*Random Forest Classifier*

*XGBoost*

*Support Vector Machine (SVM)*

*BERT (for advanced NLP)*

*Train-test split: 80-20*

*Evaluation Metrics:*

*Accuracy, Precision, Recall, F1-score*

*BERT-based model yielded the highest F1-score (~0.95), outperforming traditional ML models.*

# Visualization of Results & Model Insights

Confusion matrix to show classification performance.

ROC curves for model comparison.

Feature importance plot from Random Forest showed top TF-IDF features.

BERT attention layers showed contextual keyword dependencies.

# Tools and Technologies Used

Programming Language: Python

IDE: Jupyter Notebook, Google Colab

Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn, nltk, transformers, xgboost

Visualization: matplotlib, seaborn, plotly

# Team Members and Contributions

Mohammed Sakhee.B -Model development

Mohammed Sharuk.I-Feature Engineering

Mubarak Basha.S-EDA

Naseerudin-Data Cleaning

Rishi Kumar Baskar-Documentation and Reporting