**Phase-2 Submission**

**Student Name:** Yaswanth S  
**Register Number:** 712523121026  
**Institution:** PPG Institute of Technology  
**Department:** Bio Medical Engineering  
**Date of Submission:** 15/05/2025  
**GitHub Repository Link:** [Github](https://github.com/skyaswanth2025/NM_Fake-news-detection.git)

**1. Problem Statement**

Fake news is increasingly influencing public opinion, threatening democracy, and spreading misinformation. This project aims to build an intelligent system using Natural Language Processing (NLP) and Machine Learning (ML) techniques to classify news articles as real or fake. With a focus on binary classification, the system will analyze linguistic features and patterns in textual data to provide accurate and real-time classification.

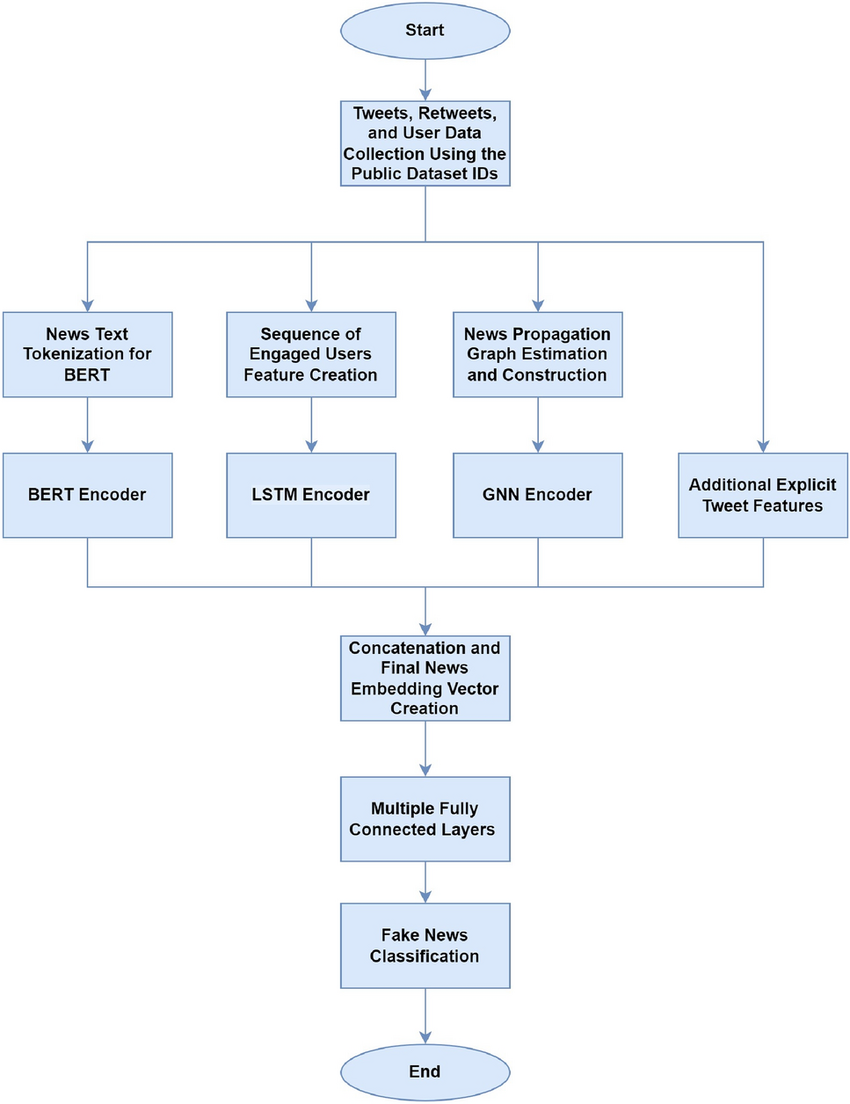
**Type of Problem:** Binary Classification (Fake vs. Real News)  
**Importance:** Enhances public awareness, prevents misinformation, and protects democratic institutions.

**2. Project Objectives**

* Apply NLP techniques for cleaning and preprocessing raw text data.
* Extract features using TF-IDF and word embeddings.
* Implement ML models (Logistic Regression, Random Forest) and DL models (LSTM, BERT).
* Evaluate models on accuracy, F1-score, and ROC AUC.
* Refine scope based on dataset insights and feedback.

**3. Flowchart of the Project Workflow**

**Data Collection → Preprocessing → EDA → Feature Engineering → Modeling → Evaluation → Deployment**



**4. Data Description**

* **Source:** Kaggle (LIAR dataset, Fake News Detection Dataset)
* **Type:** Text (Unstructured, Labeled)
* **Size:** ~12,000+ articles with attributes like author, title, content, and label
* **Nature:** Static dataset
* **Target Variable:** label (Fake/Real)

**5. Data Preprocessing**

* Removed missing/duplicate values
* Cleaned text (punctuation, numbers, special characters)
* Tokenization, stemming, and lemmatization
* Encoded text features using TF-IDF
* Normalized inputs for deep learning models

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

* Distribution of fake vs real articles
* Word frequency analysis using word clouds
* Correlation of word usage and authorship
* Key observations:
  + Fake news articles tend to use exaggerated and emotionally charged language
  + Certain sources/authors have high concentrations of fake labels

**7. Feature Engineering**

* Text-based features: TF-IDF, GloVe Embeddings
* Meta features: Title length, sentiment polarity, punctuation count
* Feature selection via correlation analysis and recursive elimination

**8. Model Building**

* Baseline: Logistic Regression, Naive Bayes
* Advanced: Random Forest, LSTM, BERT
* Data split: 80% train, 20% test (Stratified)
* Metrics used: Accuracy, Precision, Recall, F1-score

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

* Confusion matrices and ROC curves for all models
* Feature importance plot (Random Forest)
* LIME/SHAP explanations for model interpretability

**10. Tools and Technologies Used**

* **Programming Language:** Python
* **IDE/Notebook:** Google Colab, Jupyter Notebook
* **Libraries:** pandas, numpy, sklearn, matplotlib, seaborn, nltk, keras, transformers
* **Visualization Tools:** seaborn, matplotlib, WordCloud

**11. Team Members and Contributions**

| **Team Member** | **Responsibilities** |
| --- | --- |
| Yaswanth S | Problem Statement, Objectives, EDA |
| Vigneshwaran S | Data Preprocessing, Model Training |
| Suseendhiran S | Feature Engineering, Evaluation, Reporting |