**Student Name: [ k.Brundha]**

**Register Number: [ 511923106006]**

**Institution: Piyadarshini Engineering College**

**Department: Electronics And Communication Engineering**

**Date Of submission:**

**Github Repository Link**

# 1.Problem Statement

The proliferation of fake news across digital platforms poses a significant threat to public trust, political stability, and societal harmony. Traditional methods of content moderation are often slow and ineffective at scale. This project aims to develop an advanced fake news detection system leveraging Natural Language Processing (NLP) techniques to accurately identify deceptive or false information in news articles and social media posts.

# 2.Project Objectives

*Develop a machine learning model to classify news as fake or real.*

*Leverage NLP techniques for text processing and feature extraction.*

*Perform exploratory data analysis (EDA) to uncover patterns in fake vs real news.*

*Visualize model performance and insights for better interpretability.*

*Build an end-to-end pipeline for data ingestion, preprocessing, training, and evaluation.*

# 3.Flowchart of Project Workflow

**Data preprocessing**

**Feature Engineering**

**Visualization & EDA**

# 4.Data Description

**Source**: *Kaggle or Fake News Challenge datasets*

**Fields:**

**title*:*** *Title of the news article*

**text*:*** *Full text of the news content*

**subject**: *Topic category*

**label*:*** *0 for Real, 1 for Fake*

# 5.Data Processing

*Remove punctuation, HTML tags, and stopwords*

*Normalize text (lowercasing, stemming/lemmatization)*

*Tokenization*

*Handling missing values and duplicates*

# 6.Exploratory Data Analysis (EDA)

*Distribution of real vs fake news*

*Word frequency analysis (word clouds)*

*Most common words in fake vs real news*

*Text length distribution*

*Sentiment analysis (optional)*

# 7.Feature Engineering

*TF-IDF and Bag of Words (BoW)*

*N-grams (unigrams, bigrams)*

*Word embeddings (Word2Vec, GloVe, or BERT embeddings)*

*Metadata features (e.g., title length, presence of suspicious words)*

# 8.Model Building

*Machine Learning models:*

*Logistic Regression*

*Naive Bayes*

*Random Forest*

*XGBoost*

**Deep Learning**:

*LSTM / GRU*

*BERT-based Transformer*

**Evaluation Metrics:**

*Accuracy, Precision, Recall, F1-score, AUC-ROC*

# 9. Visualization of Results & Model Insights

*Confusion matrix*

*ROC Curve*

*Feature importance plots*

*Misclassification analysis*

*Word importance using LIME/SHAP*

# 10. Tools and Technologies Used

**Languages***: Python*

**Libraries:**

**NLP*:*** *NLTK, SpaCy, Transformers*

**ML/DL*:*** *scikit-learn, XGBoost, TensorFlow/Keras, PyTorch*

**Data Handling***: Pandas, NumPy*

**Visualization**: *Matplotlib, Seaborn, Plotly*

**Platforms:** *Jupyter Notebook, Google Colab, kaggle*