Phase-2

Exposing the truth with advanced fake news detection powered by natural language processing

Student Name: S. Sembaruthi

Register Number: 620123106104

Institution: AVS Engineering College

Department: Electronics and Communication Engineering

Date of Submission: 10.05.2025

Github Repository Link: https://github.com/Sembaruthi-S/Nan-Muthalvan-Project-.git

---

1.Problem Statement

In the digital age, misinformation spreads rapidly through social media and online platforms,influencing public opinion and decision-making. Detecting and mitigating the spread of fake news iscrucial for safeguarding truth and integrity in information sharing. This project aims to develop arobust NLP-powered system to classify news articles as real or fake,

aiding platforms and users inidentifying misinformation.

2.Project Objectives

To build a machine learning model that accurately detects fake news using natural languageprocessing.

- To process and analyze textual data from news sources for pattern recognition.

- To generate insights that help understand characteristics of fake news.

3.Flowchart of the Project Workflow

Data Collection -> Data Preprocessing -> Exploratory Data Analysis -> Feature Engineering ->Model Building and Evaluation -> Visualization and Insights -> Deployment

Data Collection

  - Gather news articles and headlines from reliable datasets (e.g., Kaggle, news APIs)

Data Preprocessing
   - Clean text (remove punctuation, lowercasing)
    - Tokenization
   - Stopword removal
   - Lemmatization or stemming

Exploratory Data Analysis (EDA)
   - Visualize word frequency
  - Word clouds
  - Class balance check (real vs fake)

Feature Engineering
  - TF-IDF vectorization
 - Word embeddings (e.g., Word2Vec, GloVe)
 - N-grams

Model Building
 - Train models: Logistic Regression, Random Forest, SVM, LSTM
 - Split into training and test sets

Model Evaluation
 - Evaluate with metrics: Accuracy, Precision, Recall,F1-Score
  - Confusion matrixDeployment
  - Integrate with a web application or API
  - Allow user to input a news headline/article to test authenticity

4.Data Description

 Source: Kaggle, or scraped from news websites and social media

- Features:

 - title: News title

 - text: Full news content

- subject: Topic category (e.g., politics, world news)

   - label: 1 for fake, 0 for real

5.Data Preprocessing

Removal of stop words, punctuation, and special characters Tokenization and lowercasing
  - *Lemmatization or stemming*
  - *Vectorization using TF-IDF or Word Embeddings*
  - *Handling missing/null values*

6.Exploratory Data Analysis (EDA)

Distribution of real vs fake labels:
  - *Most frequent words in fake vs real news*
  - *Word clouds*
  - *Article length distribution*
  - *N-gram analysis*

7.Feature Engineering

TF-IDF vectors

- *Count vectors*

- *Sentiment scores*

- *Word embeddings (e.g., Word2Vec, GloVe)*

- *Readability scores*

8.Model Building

Train/test split (e.g., 80/20)
 - *Classification models:*
  - *Logistic Regression*
  - *Naive Bayes*
  - *Random Forest*
  - *Support Vector Machine (SVM)*
  - *XGBoost*
  - *LSTM/GRU (Deep Learning with Keras or PyTorch)*

Evaluation metrics:
 - *Accuracy*
 - *Precision, Recall, F1-Score*
 - *Confusion Matrix*
 - *ROC-AUC Curve*

9.Visualization of Results & Model Insights

Confusion matrix heatmap

- ROC-AUC curve

- Precision-Recall curves

- Bar plots comparing model performances

- Word clouds and token frequency charts

10.Tools and Technologies Used

Languages: Python

- Libraries: Pandas, NumPy, NLTK, Scikit-learn, Matplotlib, Seaborn, TensorFlow/Keras, XGBoost

- Platforms: Jupyter Notebook, Google Colab, Kaggle

- Visualization: Matplotlib, Seaborn, WordCloud

- Version Control: GitHub

11.Team Members and Contributions

Here is a paragraph describing the roles andcontributions of Sembaruthi, Shabana, Shakthi, and Naga Ishwarya in the project "Exposing the Truth with Advanced Fake News Detection Powered by Natural Language Processing":

In this project, each team member played a vital role in ensuring a comprehensive and accurate approach to fake news detection.

S. SEMBARUTHI -data cleaning

Sembaruthi took the lead in data cleaning and preprocessing, handling the removal of noise, missing values, and formatting inconsistencies to ensure the dataset was ready for analysis.

J. Shabana Mirza - EDA and Feature Engineering

Shabana was primarily responsible for exploratory data analysis (EDA) and feature engineering—she extracted meaningful patterns, visualized trends, and transformed textual data into relevant features for the models.

G. Shakthi - Model Development

Shakthi focused on model development, training and evaluating several machine learning algorithms, including ensemble methods and deep learning models, to classify news as real or fake with high accuracy.

R. Naga Ishwarya - Documentation and Reporting

Naga Ishwarya managed documentation and reporting, compiling the project's findings, preparing visual summaries, and ensuring that all aspects of the workflow were clearly presented for stakeholders and final submission.