Phase-1

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 Communications Engineering

Date of Submission: 30.04.2025

1.Problem Statement

the digital age, misinformation and fake news have become increasingly prevalent, affecting public opinion, political discourse, and societal trust. Traditional methods of detecting misinformation are no longer sufficient due to the sheer volume and sophistication of fake content. There is a pressing need for an automated, intelligent system capable of accurately identifying and flagging fake news articles using modern Natural Language Processing (NLP) techniques.

2.Objectives of the Project

Develop a robust NLP-based system to detect fake news from textual content with high accuracy.

Classify news articles as real or fake using machine learning and deep learning models.

Provide insights and confidence scores to enhance user trust and interpretability.

Create a user-friendly interface to demonstrate real-time fake news detection.Evaluate model performance using industry-standard metrics and optimize for real-world deployment.

3.Scope of the Project

In Scope:

Text-based news content (articles, social media posts, etc.).

Supervised learning models including traditional ML

and deep learning-based NLP models.

English language content.

Dashboard for fake news classification and analysis.

Out of Scope:

Multimedia content (e.g., image or video-based fake news).

Non-English language processing.

Legal or forensic applications of misinformation analysis.

4.Data Sources

Datasets:

LIAR Dataset: Contains labeled short statements from Politifact.

Fake and Real News Dataset from Kaggle: Contains fake and real news articles with labeled sources.

BuzzFeed and Politifact articles: Used for further validation.

Preprocessing Tasks:

Text cleaning (stopwords removal, lemmatization).

Tokenization and vectorization (TF-IDF, Word2Vec, BERT embeddings).

5.High-Level Methodology

1.Data Collection & Preprocessing: Import, clean, and normalize data.

2.Exploratory Data Analysis (EDA): Visualize word usage, length distributions, and source credibility.

3.Feature Engineering: Use NLP techniques for feature extraction.

Model Development:

Baseline ML models: Logistic Regression, Random Forest, SVM.

Advanced models: LSTM, BERT, RoBERTa.

Evaluation: Accuracy, Precision, Recall, F1-score, ROC-AUC.

6.Tools and Technologies

Programming Languages & Libraries

Python – Primary programming language for model development and data processing.

Pandas & NumPy – Data manipulation and numerical operations.

Scikit-learn – Traditional machine learning models and utilities.

NLTK / SpaCy – NLP preprocessing (tokenization,, lemmatization, etc.).

TensorFlow / PyTorch – Deep learning model development (e.g., LSTM, BERT).

Transformers (Hugging Face) – For implementing pre-trained models like BERT, RoBERTa.

Matplotlib / Seaborn / Plotly – Data visualization for EDA and results.

Modeling Techniques

TF-IDF, Word2Vec, BERT Embeddings – Text vectorization and semantic analysis.

Logistic Regression, Random Forest, SVM – Baseline ML classifiers.

LSTM, GRU – Sequential deep learning models.

BERT, RoBERTa – Transformer-based models for state-of-the-art text classification.

Development Tools

Jupyter Notebooks / VS Code – For coding and experimentation.

Git / GitHub – Version control and collaboration.

Docker – Containerization for reproducibility and deployment.

Streamlit / Flask – For building interactive web-based user interfaces.

Postman – API testing.

Deployment Platforms

Heroku / Render / AWS / GCP – For deploying the application to the cloud.

7.Team Members and Roles

S. Sembaruthi –Project Lead & Data Scientist

Oversees the overall project development and timeline.Designs and implementsmachine learning models.Performs data preprocessing, feature engineering, andmodel evaluation..Ensures model explainability, accuracy, and compliance withhealthcare standards.

G.Shakthi –Software Developer & System Integrator

Develops the front-end and back-end for the prediction system (web or mobileinterface).Integrates the trained AI model into the application.Managesdeployment using cloud platforms (e.g., AWS, Heroku).Ensures UI/UX is userfriendly for healthcare professionals.

J. Shabana Mirza and R. Naga Ishwariya-Data Engineer & Analyst

Collects, cleans, and prepares patient datasets from various sources.Handlesdatabase management and data pipelines.Conducts exploratory data analysis(EDA) and generates visual insights.Works with the data scientist to ensure high-quality training data.