Fake News Detection using NLP and Machine Learning

Objective

Build a machine learning system that classifies news headlines or articles as real or fake using natural language processing (NLP) techniques and classification algorithms.

Tech Stack

Languages & Libraries:

- Python
- scikit-learn, TensorFlow, Hugging Face Transformers

NLP Tools:

- NLTK, spaCy

Datasets:

- LIAR dataset
- ISOT dataset
- Other fake news datasets from Kaggle

Development Phases

1. Data Collection (completed)

- Use publicly available labeled datasets with real and fake news samples
- Ensure a balanced distribution of both classes

2. Text Preprocessing (completed)

- Clean the news text:
    - Remove stopwords
    - Apply lemmatization or stemming
    - Convert to lowercase, remove punctuation
- Vectorize text using:
    - TF-IDF (Term Frequency–Inverse Document Frequency)
    - Word embeddings like GloVe or BERT

3. Model Building (completed)

- Train and test different models such as:
    - Logistic Regression
    - Random Forest
    - Support Vector Machine (SVM)
    - Deep learning model using BERT or LSTM

4. Evaluation (Completed)

- Evaluate model performance using metrics:
    - Accuracy
    - Precision
    - Recall
    - F1-Score
- Use confusion matrix and ROC curves for deeper insights

5. Frontend

- Develop a simple web app using Flask
    - Input: User enters a headline or paragraph
    - Output: Model returns prediction as “Real” or “Fake”

Expected Outcomes

- A working fake news classifier with good accuracy and reliability
- Understanding of how NLP techniques contribute to detecting misinformation
- Potential for integration into social media platforms or news aggregators