Fake News Detection Using NLP

# 1. Introduction

Fake news has become a significant problem in today’s digital world. This project aims to detect fake news using natural language processing (NLP) techniques and machine learning models. The goal is to classify news articles as real or fake by analyzing their content.

# 2. Dataset Description

The dataset used for this project is the Kaggle Fake News Dataset, which contains news articles that are labeled as either fake or real. The dataset consists of the following features: Title, Text, and Label (1 for fake news, 0 for real news).

# 3. Data Preprocessing

The preprocessing steps include:  
- Tokenization: Splitting the text into individual tokens (words).  
- Stop-word Removal: Removing common words like 'the', 'is', etc.  
- Lemmatization: Reducing words to their base forms (e.g., 'running' -> 'run').  
- Vectorization: Converting text data into numerical format using techniques like TF-IDF and Bag of Words.

# 4. Exploratory Data Analysis (EDA)

EDA involved visualizing the frequency of words in real and fake news articles using word clouds, identifying patterns and keywords that were more common in fake news. Various visualizations helped us understand the data better.

# 5. Model Building

Several machine learning algorithms were used to build the model, including Logistic Regression, Naive Bayes, Support Vector Machine (SVM), and Random Forest. TF-IDF was used for feature extraction, and the models were trained and evaluated using the dataset.

# 6. Results and Evaluation

The models were evaluated using performance metrics such as Accuracy, Precision, Recall, and F1 Score. The confusion matrix was also used to analyze the performance of the models in predicting real vs fake news.

# 7. Conclusion

The project successfully demonstrated that fake news detection can be achieved using NLP techniques and machine learning models. The best-performing model was able to classify news articles with a high degree of accuracy. Future work could include experimenting with more complex models and larger datasets to improve performance.