📝 Project 6: Fake News Article Classification

➢ Introduction

➢ In today's digital age, misinformation is one of the biggest threats to reliable communication. Fake news can spread rapidly and influence public opinion, elections, and even health decisions. This project aims to create a machine learning model that classifies news articles as either fake or real using Natural Language Processing (NLP) techniques. The goal is to contribute to automated detection systems that can help combat the spread of false information.

➢ Abstract

➢ This project focuses on developing a binary classification model that distinguishes fake news from real news using textual data. The dataset comprises real news articles from verified publishers and fake news articles from various unverified sources. We perform preprocessing on the text including lowercasing, stopword removal, punctuation removal, and stemming. We then transform the text into numerical features using TF-IDF Vectorizer and train two machine learning models: **Logistic Regression** and **Multinomial Naive Bayes**. The performance of both models is evaluated using accuracy and classification reports. The best-performing model is saved and can be used for future predictions or deployment.

➢ Tools Used

➢ **Programming Language:** Python (executed on Google Colab)

➢ **Libraries:**

- pandas – for data handling
- nltk – for text preprocessing
- scikit-learn – for TF-IDF, model training, and evaluation
- joblib – for model saving

- matplotlib, seaborn – for visualization

- **Dataset Source:** Kaggle (Files: Fake.csv, True.csv)
- **Techniques Used:** TF-IDF Vectorization, Logistic Regression, Naive Bayes

Steps Involved in Building the Project

- **Data Collection:**
- **Data Merging:**
- **Text Cleaning:**
- **Feature Extraction:**
- **Train/Test Split:**
- **Model Training:**
- **Model Evaluation:**
- **Model Saving:**

Conclusion

- This project demonstrates how machine learning and NLP can work together to automate the task of fake news detection. Both Logistic Regression and Naive Bayes models performed well, with Naive Bayes showing slightly higher accuracy for this dataset. The trained model and vectorizer were saved and can now be integrated into applications or APIs to help prevent the spread of misinformation on the internet.

- **Author:** shrey sakhiya
  Date: 26/7/2025