📰 Fake News Detection Using Machine Learning

Abstract

Fake news spreads rapidly through digital platforms and can mislead people and affect society. This project focuses on building a Fake News Detection system using Machine Learning and Natural Language Processing techniques. A labeled dataset from Kaggle was used to train the model. TF-IDF vectorization and Logistic Regression were applied to classify news articles as Fake or Real. A Streamlit web application was developed to allow users to check news authenticity in real time

Introduction

With the increase in online news consumption, the problem of fake news has become more serious. Manual verification of news articles is difficult due to the large volume of information. Machine Learning provides an automated way to analyze and classify news content. This project aims to design a system that can help users identify fake and real news using text classification techniques.

Tools Used

The following tools and technologies were used to build this project:

- **Programming Language:** Python
- **Libraries:**
  - Pandas and NumPy for data processing
  - Scikit-learn for machine learning and evaluation
  - NLTK for text preprocessing
- **Feature Extraction:** TF-IDF Vectorizer
- **Machine Learning Model:** Logistic Regression
- **Development Environment:** Jupyter Notebook
- **Web Application Framework:** Streamlit
- **Model Storage:** Pickle
- **Dataset Source:** Kaggle (Fake and Real News Dataset)

Steps Involved in Building the Project

1. **Dataset Collection:**
   A labeled dataset containing fake and real news articles was collected from Kaggle. The dataset consisted of two separate files for Fake and Real news.

2. **Data Preprocessing:**
   Text data was cleaned by converting it to lowercase, removing special characters, URLs, and unnecessary symbols. Stopwords were removed to improve text quality and reduce noise.

3. **Label Encoding:**
   Fake news articles were labeled as 0 and real news articles as 1. Both datasets were merged and shuffled to create a balanced dataset.

4. **Feature Extraction:**
   TF-IDF (Term Frequency–Inverse Document Frequency) vectorization was applied to transform text data into numerical features that can be used by machine learning algorithms.

5. **Model Training:**
   A Logistic Regression classifier was trained using the processed dataset with a train-test split method.

6. **Model Evaluation:**
   The trained model was evaluated using accuracy, precision, recall, and F1-score. The model achieved an accuracy of approximately 99%, showing strong classification performance.

7. **Model Saving:**
   The trained model and vectorizer were saved using Pickle for reuse in the web application.

8. **Web Application Development:**
   A Streamlit-based web application was created to allow users to enter news text and receive predictions along with probability scores and key word signals.

Conclusion

The Fake News Detection system successfully classifies news articles as Fake or Real with high accuracy. The Streamlit application makes the model easy to use for real-time predictions. Although the model performs well on the dataset, it can be improved in the future by training on more diverse news sources. This project demonstrates the importance of Machine Learning in reducing misinformation and supporting trustworthy news consumption.