



News Article Classification using Machine Learning

♦ Introduction

Fake news has become a major challenge in today's digital world. Misinformation can spread rapidly, impacting public opinion, politics, and health. This project aims to classify news articles as either **FAKE** or **REAL** using natural language processing (NLP) and machine learning (ML) techniques. By automating the classification process, this model can assist users, journalists, and organizations in identifying misleading content.

♦ Abstract

The goal of this project is to build a machine learning-based classifier that predicts whether a given news article is fake or real. We used a labeled dataset from Kaggle consisting of real news from verified sources and fake news articles collected from unreliable websites. The model uses TF-IDF vectorization for feature extraction and Logistic Regression for classification. A user-friendly Streamlit web application allows users to input custom news text and receive a prediction in real time.

♦ Tools & Technologies Used

- **Python 3**
 - **Pandas** – Data handling
 - **Scikit-learn** – Model training and evaluation
 - **TfidfVectorizer** – Feature extraction from text
 - **Logistic Regression** – Classification algorithm
 - **Streamlit** – Web app UI for user interaction
 - **Jupyter Notebook** – Model development and testing environment
-

♦ Steps Involved in the Project

1. Dataset Collection

The dataset was sourced from Kaggle. It contains two CSV files: **Fake.csv** and **True.csv**.

2. Data Preparation

The two datasets were merged into a single file, `news_kaggle_merged.csv`, and labeled appropriately (FAKE = 0, REAL = 1). Irrelevant columns were dropped to keep only the text content and labels.

3. Text Preprocessing

Text data was cleaned and vectorized using `TfidfVectorizer`, which converts raw text into numerical features while removing stopwords and common terms.

4. Model Training

A **Logistic Regression** model was trained using an 80/20 train-test split. The model was able to learn patterns distinguishing real news from fake.

5. Evaluation

The classifier was evaluated using standard metrics like accuracy, precision, recall, and F1-score. Performance was strong on the test set.

6. Web Application

A **Streamlit** web interface was built to make the model accessible to users. Users can input news text and get instant predictions.

◆ Conclusion

This project demonstrates how NLP and machine learning can be used to combat misinformation by detecting fake news. With minimal resources, it's possible to create a reliable text classifier and deploy it using simple tools like Streamlit. The model can be further enhanced with deep learning, more complex NLP preprocessing, or additional metadata (e.g., authorship, domain). Overall, the project provides a practical and effective solution for real-world fake news detection.