News Article Classification (Real/Fake) Using Machine Learning and NLP

Introduction

Fake news is a growing concern in the digital era, affecting public opinion and trust in media. This project addresses the challenge by leveraging Natural Language Processing (NLP) and Machine Learning (ML) to build an automated news classification system.

Abstract

The objective of this project is to develop a model capable of classifying news articles as real or fake. Using NLP and ML techniques, the system detects patterns in textual data to identify misinformation. The system aims to ensure accuracy, reliability, and scalability for real-time deployment. Through a combination of preprocessing, vectorization, and model evaluation, the project demonstrates high classification performance.

Tools Used

- Programming Language: Python
- Data Manipulation: Pandas, NumPy
- Machine Learning: Scikit-learn (Logistic Regression, Naïve Bayes)
- Text Processing: NLTK (tokenization, stopword removal, lemmatization)
- Visualization: Matplotlib, Seaborn
- Development: Jupyter Notebook

Steps Involved in Building the Project

1. Data Collection: Gathered real and fake news articles into a balanced dataset.
2. Data Preprocessing: Cleaned text (lowercasing, punctuation removal, lemmatization).
3. Feature Extraction: Applied TF-IDF vectorization to convert text into numerical form.
4. Model Training: Trained models using Logistic Regression and Naïve Bayes.
5. Evaluation: Measured performance using accuracy, precision, recall, F1-score, and confusion matrix.
6. Deployment: Built a simple interface to test real-time predictions.

Conclusion

The fake news detection system was successfully implemented using NLP and ML. The Logistic Regression model achieved 92% accuracy, and Naïve Bayes achieved 89%. TF-IDF vectorization provided better performance than Bag-of-Words. The project demonstrates effective classification of news articles, with potential for real-time deployment. Future work can include integration of deep learning models and larger datasets.