News Article Classification Using Machine Learning

Report by Abhinav R

Introduction

In the digital era, the uncontrolled spread of misinformation through online platforms has become a serious issue. Fake news articles—crafted to mislead readers—can manipulate public opinion, influence elections, and cause panic during emergencies. As part of my internship with Elevate Labs, I developed a machine learning-based news classification system to detect fake news based on its text content. This project combines Natural Language Processing (NLP) and machine learning techniques to build an automated, scalable solution to identify misinformation.

Abstract

The primary goal of this project was to classify news articles as either real or fake using supervised machine learning. I utilized publicly available datasets containing labeled real and fake news. The raw text data was cleaned and transformed into meaningful numerical features using TF-IDF vectorization. A Logistic Regression model was trained and evaluated, achieving impressive performance metrics. The final model was deployed using Streamlit, offering users an interactive platform to test the model by entering news article text.

Tools and Technologies Used

- Programming Language: Python
- Libraries: Pandas, NumPy, Scikit-learn, NLTK, Joblib
- IDE: Jupyter Notebook
- Deployment: Streamlit
- Version Control: GitHub

Steps Involved in Building the Project

1. Data Collection:
   Two datasets (True.csv and Fake.csv) were collected. Each contained thousands of news articles. A new column labeled each article as 1 (real) or 0 (fake). These datasets were merged into one for further processing.
2. Preprocessing:
   Text preprocessing was crucial. It involved:

- Lowercasing all words
- Removing punctuation, special characters, and digits
- Removing stopwords using NLTK
- Tokenizing the text into meaningful words

3. Vectorization:
   Cleaned text was converted into numerical format using the TF-IDF (Term Frequency-Inverse Document Frequency) method. This allowed the model to learn which words were more informative for classification.
4. Model Training:
   A Logistic Regression model was chosen for its simplicity and effectiveness in binary classification tasks. The dataset was split into 80% training and 20% testing. The model was trained on the vectorized text and evaluated for accuracy.
5. Model Evaluation:
   The model achieved high accuracy (~95%). Evaluation metrics included:
   - Accuracy Score
   - Precision, Recall, F1-Score
   - Confusion Matrix
     This showed that the model performed well across both classes, with minimal false positives and false negatives.
6. Deployment:
   A front-end interface was developed using Streamlit, where users can input a news article. The application returns whether the article is real or fake in real time. The trained model was serialized using joblib and loaded into the app.

Conclusion

This project highlights the potential of machine learning and NLP in real-world applications like fake news detection. The combination of preprocessing, TF-IDF vectorization, and logistic regression resulted in a highly accurate classifier. Moreover, deploying it through a web interface makes it accessible and user-friendly.