News Article Classification (Fake/Real)

Title:

News Article Classification (Fake/Real)

Objective:

To build a machine learning model that classifies news articles as Fake or Real using Natural Language Processing (NLP) techniques.

1. Introduction

Fake news has become a major concern with the rise of digital media. Misinformation can spread rapidly and cause real-world consequences. This project aims to develop a classifier that can detect fake news using text-based features extracted from article content.

2. Tools and Libraries Used

- Python
- Pandas
- NumPy
- NLTK
- Scikit-learn (sklearn)
- TF-IDF Vectorizer
- Logistic Regression
- Streamlit (optional)

3. Dataset Description

The dataset contains labeled news articles. Each article is categorized as either:

- Real: Authentic and verified information
- Fake: Misleading or fabricated content

Columns used:

- text: Full content of the news article

- label: Binary value (0 = Fake, 1 = Real)

4. Data Preprocessing Steps

4.1. Handling Missing Data: Dropped null entries

4.2. Text Cleaning (Using NLTK):

- Lowercasing

- Removing punctuation and special characters

- Removing stopwords

- Lemmatization

5. Feature Extraction

Used TF-IDF Vectorization to convert textual data into numerical features:

TfidfVectorizer(max_features=5000, stop_words='english')

6. Model Implementation

6.1. Data Split: train_test_split(X, y, test_size=0.2, random_state=42)

6.2. Algorithm Used: Logistic Regression

7. Evaluation Metrics

Accuracy: 97.73%

Precision: 97.74%

Recall: 97.73%

F1 Score: 97.72%

8. Output Sample

Fig 1. Sample Model Prediction Output

The model correctly classifies unseen text into 'Real' or 'Fake' news.

9. Conclusion

- The logistic regression model achieved 97%+ accuracy using TF-IDF features.

- Text preprocessing using NLTK and vectorization was critical for performance.

- The model can be deployed using Streamlit for real-time prediction.

10. Future Work

- Use more advanced models like BERT or LSTM

- Include title/author metadata as features

- Develop a feedback loop to improve accuracy over time.