Fake News Detection Using Machine Learning

Introduction

The spread of misinformation in digital media makes fake news detection crucial. This project develops a machine learning model to classify news articles as real or fake using **Naïve Bayes, Random Forest, and LSTM**.

Dataset

The **Fake and Real News Dataset** contains labeled news articles with:

- **Title**: Headline
- **Text**: Full content
- **Label**: Real or Fake

Data Preprocessing

Key preprocessing steps:

- **Stopword Removal**: Eliminates common words.
- **Stemming & Lemmatization**: Converts words to root forms.
- **Tokenization**: Splits text into words/phrases.
- **TF-IDF Vectorization**: Converts text into numerical data.

Model Training

1. Naïve Bayes

- Probabilistic model based on Bayes' Theorem.
- Efficient for word frequency analysis.

2. Random Forest

- Ensemble method using multiple decision trees.
- High accuracy, prevents overfitting.

3. LSTM

- Deep learning model for sequential data.
- Captures contextual relationships in text.

Evaluation Metrics

Models are assessed using:

- **Accuracy**: Overall correctness.
- **Precision**: Correctly identified fake news percentage.
- **Recall**: Ability to detect all fake news.
- **F1-Score**: Balance between precision and recall.

Results & Insights

- **Naïve Bayes**: Fast but struggles with complex patterns.
- **Random Forest**: More accurate but needs feature engineering.
- **LSTM**: Best at contextual understanding but resource-intensive.

Deployment

A **Flask web app** allows users to input news articles and get real/fake predictions, with **HTML, CSS, and JavaScript** for the frontend.

Conclusion

The system effectively identifies fake news. Future improvements include **BERT-based NLP models** and dataset expansion for better accuracy.