Project Report: News Article Classification (Fake or Real)

1. Introduction

Fake news has become a critical issue in today's digital world, with misinformation spreading rapidly through social media and other online platforms. The objective of this project is to develop a machine learning model to classify news articles as either **fake** or **real** using Natural Language Processing (NLP) techniques. A user-friendly web interface is created using Streamlit to allow real-time classification of news articles.

2. Objectives

- Build a reliable classification model using labeled news datasets.
- Apply text preprocessing techniques to clean and prepare the data.
- Vectorize text data using TF-IDF.
- Train and evaluate machine learning models such as Logistic Regression and Naive Bayes.
- Develop a web-based interface via Streamlit to interactively classify news articles.

3. Tools and Technologies

- **Programming Language:** Python
- **Libraries:** Scikit-learn, Pandas, NLTK, Joblib, Streamlit
- **Dataset:** Labeled fake and real news datasets from Kaggle
- **Development Environment:** Jupyter Notebook

4. Methodology

4.1 Data Collection

A publicly available dataset containing labeled news articles was obtained from Kaggle. The dataset includes two classes: fake news and real news articles.

4.2 Data Preprocessing

Text data was cleaned using the NLTK library including:

- Removing stopwords, punctuation, and special characters
- Lowercasing all text

- Tokenization
- Stemming / Lemmatization

This prepares the data for effective learning by the model.

4.3 Feature Extraction

TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer was used to transform the cleaned text into numerical vectors, capturing both word importance and frequency while reducing the weight of commonly used words.

4.4 Model Training

Two classification algorithms were trained on the vectorized data:

- Logistic Regression
- Naive Bayes (MultinomialNB)

The dataset was split into training and testing sets for model evaluation.

4.5 Evaluation

Models were evaluated with the following metrics:

- Accuracy
- Precision
- Recall
- F1 Score

Among the two, the model with better performance was selected for deployment.

4.6 Deployment

A Streamlit web app was developed to provide a simple, interactive interface where users can input news articles and receive classification results instantly. The app also provides explanation and confidence scores.

5. Results

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	0.92	0.93	0.90	0.91

Model	Accuracy	Precision	Recall	F1 Score
Naive Bayes	0.89	0.90	0.87	0.88

Logistic Regression showed higher accuracy and F1 score and was selected for the web app.

6. Deliverables

- Jupyter notebook containing code for data cleaning, feature extraction, model training, and evaluation
- Trained model and vectorizer saved using Joblib (lr_model.joblib, vectorizer.joblib)
- Streamlit web app (app.py) for live classification
- This project report document

7. Conclusion

The project successfully demonstrates the power of NLP and machine learning in detecting fake news articles. The end-to-end pipeline—from data preprocessing, modeling, to web app deployment—enables easy verification of news authenticity in real time. Such tools can help combat misinformation by raising awareness and providing quick fact checks.

8. Future Work

- Enhance the model with more sophisticated deep learning models such as transformers
- Incorporate explainable AI techniques to highlight which parts of the text influenced the prediction
- Deploy the app on cloud platforms for public access

9. References

- Kaggle Datasets — Fake News Detection
- Scikit-learn Documentation
- NLTK Library Documentation
- Streamlit Documentation

Prepared by: [Yashit Chugh]
Date: [29/7/25]