NAME – Kumud Metkar

Fake News Detection Using NLP

1–2 Page Project Report

1. Introduction

With the rapid spread of information on the internet, detecting fake news has become an important problem. The aim of this project is to build a machine learning model that classifies news articles as **real or fake** based on their content using **Natural Language Processing (NLP)** techniques.

---

2. Abstract

This project focuses on detecting fake news by applying machine learning on a labeled dataset of news articles. The dataset was collected from Kaggle, containing both fake and real news samples. The main goal is to process the text using NLP, train a classification model, and provide a live web-based interface for prediction using Streamlit. The final application allows users to paste news content and see whether it is likely fake or real.

---

3. Tools Used

- **Programming Language**: Python
- **Libraries**: Pandas, NLTK, scikit-learn, Streamlit, Pickle
- **IDE**: Jupyter Notebook
- **Web Interface**: Streamlit
- **Dataset Source**: Kaggle (Fake and Real News Dataset)

---

4. Steps Involved in Building the Project

1. **Data Collection**:
   Fake and real news datasets were downloaded from Kaggle.
2. **Data Preprocessing**:
   - Combined fake and real datasets.
   - Cleaned text using NLTK (stopword removal, stemming, punctuation removal).
   - Created new "text" field by merging title and article content.

3. **Feature Extraction**:
    - Applied **TF-IDF Vectorization** to convert text to numerical features.
4. **Model Training**:
    - Used **Logistic Regression** for classification.
    - Evaluated the model using Accuracy and F1-Score.
5. **Model Saving**:
    - Saved the trained model and vectorizer using pickle.
6. **Web Interface**:
    - Built a user-friendly Streamlit app where users can input news text and get instant predictions.

---

5. Conclusion

The project successfully demonstrates the use of NLP and machine learning in identifying fake news articles. With a simple interface, it allows users to evaluate the credibility of news content. This solution can be further enhanced by using deep learning models and larger datasets for improved accuracy.