Fake News Detection Project Report

Introduction

In the modern era, the rapid spread of fake news on social media and online platforms can cause significant social, political, and financial impacts. Detecting fake news automatically has become an essential task. This project aims to build a machine learning-based system to classify news articles as **Fake** or **Real** using Natural Language Processing techniques.

Abstract

Fake news detection is a critical problem due to the widespread dissemination of misinformation. This project combines text preprocessing, feature extraction using TF-IDF, and a Logistic Regression classifier to build a robust fake news detection system. The model is trained on labeled datasets containing fake and real news articles, achieving high accuracy in distinguishing genuine content from misleading content.

Tools Used

- **Programming Language:** Python
- **Libraries:** pandas, scikit-learn, re, joblib
- **Data Handling:** CSV files (Fake.csv and True.csv)
- **Machine Learning:** Logistic Regression, TF-IDF Vectorizer
- **Environment:** Jupyter Notebook / VS Code

Steps Involved in Building the Project

1. **Dataset Collection:**
    - Combined Fake.csv and True.csv datasets with a label column (0 = Fake, 1 = Real).
2. **Data Preprocessing:**
    - Cleaned text by removing punctuations, numbers, and special characters.
    - Converted text to lowercase for uniformity.
3. **Data Splitting:**
    - Split data into training (75%) and testing (25%) sets.
4. **Feature Extraction:**

- Used **TF-IDF Vectorizer** to convert text into numerical features suitable for machine learning.

5. **Model Training:**
   - Trained **Logistic Regression** on the TF-IDF features of the training set.
6. **Evaluation:**
   - Evaluated the model using Accuracy, Precision, Recall, and F1-Score.
   - Achieved high performance in detecting fake news articles.
7. **Model Deployment (Optional):**
   - Saved trained model using joblib for future use in applications.

Conclusion

This project successfully builds a **fake news detection system** using Python and machine learning techniques. The Logistic Regression model, combined with TF-IDF features, provides a reliable way to distinguish between fake and real news with high accuracy. This system can serve as a practical tool for users and organizations to filter misinformation online.

🔸 **Notes:**

- The report is concise and fits within 2 pages when formatted in standard PDF style (A4, 11–12pt font, single spacing).
- Optional visualizations (like accuracy charts) can be added if needed but are not required for the 2-page limit.