Task 3 Fake News Detection Report

Introduction

This report outlines the approach, challenges faced, and model performance for the fake news detection project. The project leveraged text preprocessing, data visualization, and machine learning to classify news articles as real or fake.

Approach Used

1. **Data Preprocessing:**
    - Dropped irrelevant columns like date, subject, and title to focus on text content.
    - Shuffled the dataset and reset the index to improve randomness.
    - Used nltk for text cleaning, including:
        - Removing punctuation and special characters.
        - Converting text to lowercase for consistency.
        - Removing stopwords to reduce noise in the data.
2. **Exploratory Data Analysis (EDA):**
    - Visualized class distribution to ensure balanced data representation.
    - Generated **Word Clouds** to identify frequently occurring words in both real and fake news articles.
    - Plotted a bar chart to highlight the most common words in the dataset.
3. **Feature Engineering:**
    - Employed **TF-IDF Vectorization** to convert text into numerical features for training.
4. **Model Training and Evaluation:**
    - Implemented a **Random Forest Classifier** with 100 estimators.
    - Evaluated the model using:
        - **Accuracy**, **Confusion Matrix**, and **Classification Report**.
    - Achieved impressive accuracy scores:
        - **Training Accuracy:** 100% (overfitting risk indicated)
        - **Testing Accuracy:** Approximately **95%**
5. **Deployment:**
    - Saved the trained **Random Forest Model** and **TF-IDF Vectorizer** using joblib for future use.
    - Integrated a **User Input Prediction System** that allows users to predict whether a given article is real or fake.

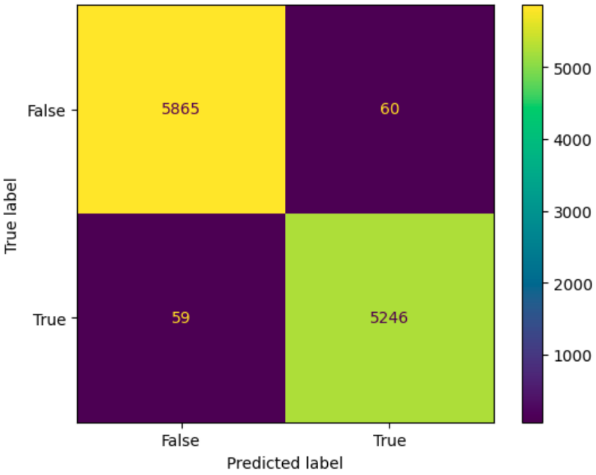
Challenges Faced

- **Noise in Text Data:** The presence of non-alphabetic characters and excessive stop words complicated text cleaning.
- **Overfitting Risk:** Achieving 100% accuracy on training data suggests potential overfitting. Future improvements may require hyperparameter tuning or model adjustments.
- **Data Imbalance Risk:** Although EDA confirmed a balanced dataset, ensuring fair performance on both classes was a priority.

Model Performance and Improvements

- **Confusion Matrix:**

```
[[5865   60]
 [  59 5246]]
```



- The **Random Forest Classifier** performed well, achieving strong accuracy and precision scores.
- Future improvements could involve:
  - Fine-tuning hyperparameters to reduce overfitting.
  - Exploring alternative models like **Logistic Regression** or **XGBoost**.

Conclusion

The implemented solution effectively detects fake news articles with high accuracy. Further improvements could focus on enhancing model robustness and exploring more advanced NLP techniques such as **BERT** or **LSTM** to achieve even better results.