Fake News Detection with Sentiment analysis & ANN

Overview

This project aimed to develop a system for classifying news articles into real or fake categories using Artificial Neural Networks (ANN) and sentiment analysis. It involved data collection, preprocessing, and feature engineering to prepare the text data for analysis. The ANN model was trained on the preprocessed data to identify complex patterns, while sentiment analysis provided insight into the emotional context of the articles. The model's performance was evaluated, and a user-friendly web interface was created using Streamlit for real-time predictions. Through this integration, the project aimed to combat the spread of misinformation effectively.

About Dataset

The dataset was taken from The Online Academic Community, University of Victoria. The ISOT Fake News dataset is a compilation of several thousands fake news and truthful articles, obtained from different legitimate news sites and sites flagged as unreliable by Politifact.com.

The dataset contains different types of articles on different topics, however, the majority of articles focus on political and World news topics. The dataset consists of two CSV files. The first file named "True.csv" contains more than 12,600 articles from reuter.com. The second file named "Fake.csv" contains more than 12,600 articles from different fake news outlet resources. Each article contains the following information: article title, text, type and the date the article was published on. To match the fake news data collected for kaggle.com, we focused mostly on collecting articles from 2016 to 2017.

Methodologies

- **Data Preprocessing :**
    - Since there was two sets of data first goal was to merge them into single dataset with a flag('target') indicating the text is true(0) or fake(0).
    - Dropping Unnecessary columns – title and date
    - There were no null values
    - Lastly the data is shuffled to doesn't induce bias the model's learning process.

- **Text Preprocessing :**
    - Tokenization using ***TweetTokenizer***
    - Removing Special Characters by ***Regular Expression***
    - Lemmatizing using ***WordNetLemmatizer***

  - Converting to ***Lowercase characters***
  - Removing ***Stopwords***
  - Transform the text into meaningful representation of integers or numbers using **TF-IDF Vectorizer**

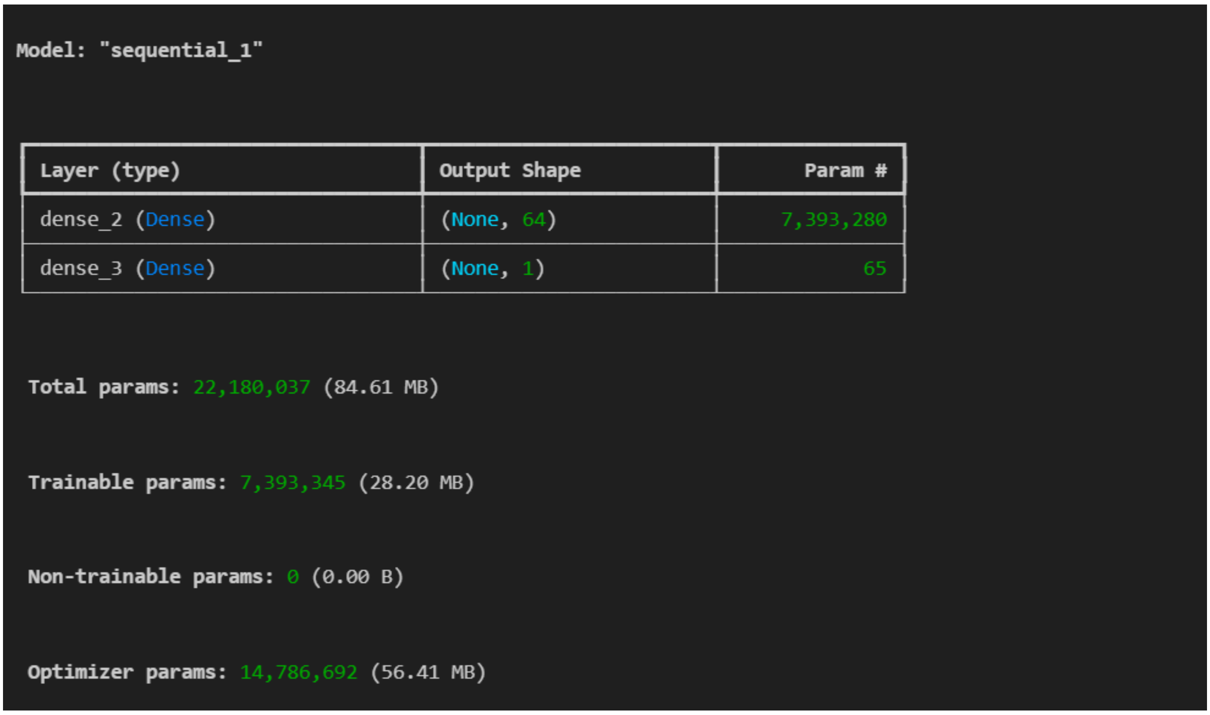
- **Splitting data into train and test set (70-30)**
- **Scaling Training data :**

  - Scaling training data involves transforming the features to a similar scale, which is crucial for many machine learning algorithms to perform effectively. MaxAbsScaler is used in this project because it scales each feature by its maximum absolute value, preserving the data's sparsity and ensuring that all features fall within the range [-1, 1]. This scaler is particularly useful for text data, like that used in fake news detection, as it maintains the interpretability of the features while preventing outliers from dominating the learning process.

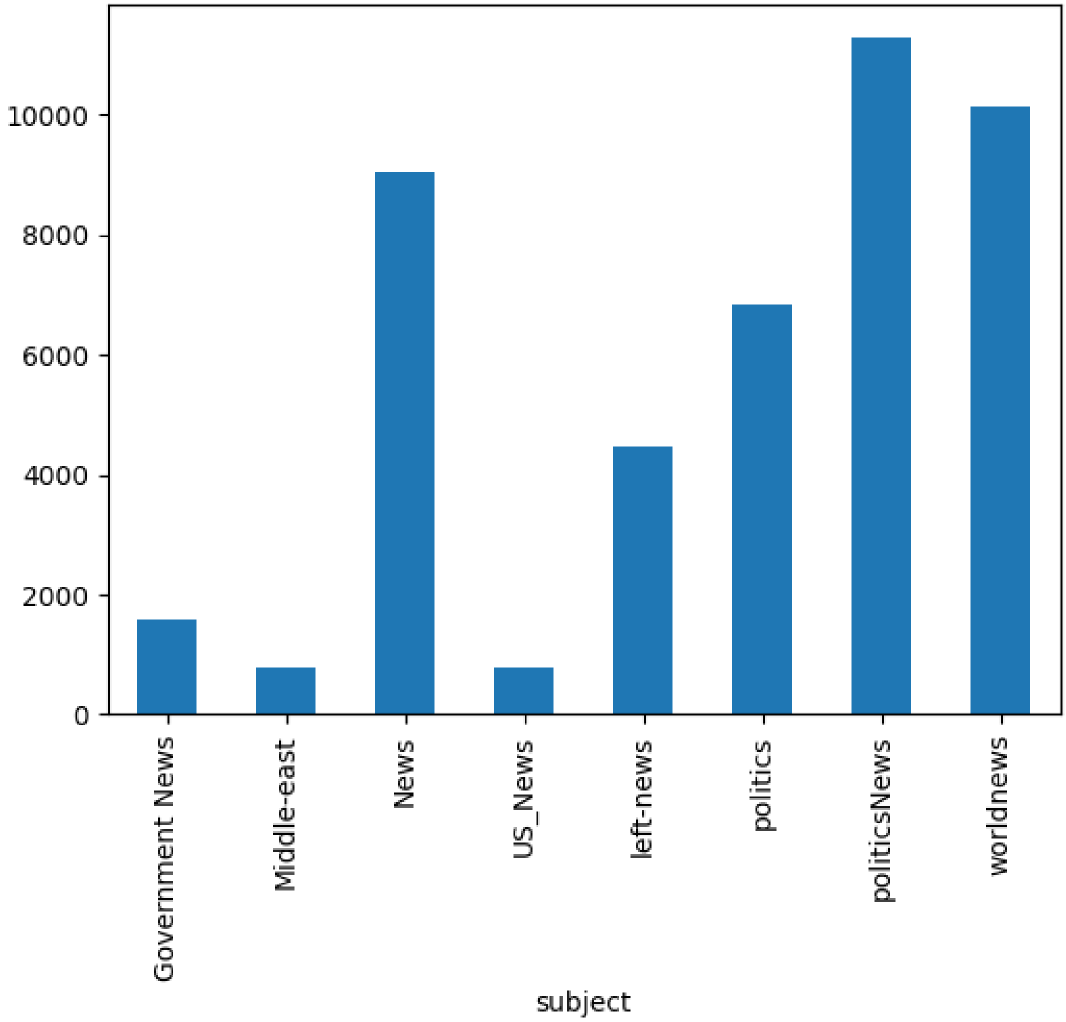
- **Setting Up ANN Model :**

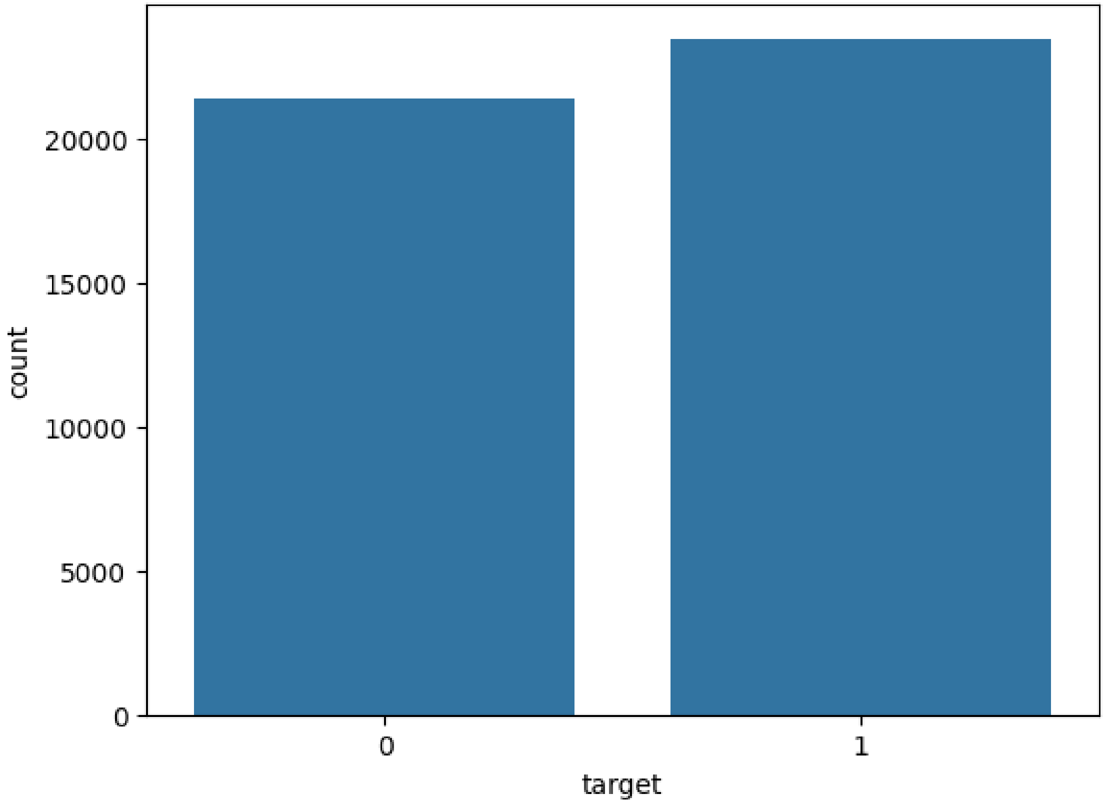
- The model architecture is Sequential, meaning layers are added sequentially.
- It consists of two dense layers:
  - The first dense layer has 64 neurons.
  - The second dense layer has a single neuron.
- The activation function for the first layer is Rectified Linear Unit (ReLU), which introduces non-linearity to the model.
- The activation function for the second layer is Sigmoid, suitable for binary classification tasks as it squashes the output between 0 and 1.
- The input shape for the first layer is determined by the shape of the scaled training data.
- The model is compiled using binary cross-entropy as the loss function, which is commonly used for binary classification tasks.
- The Adam optimizer is used for optimizing the model parameters during training.
- Accuracy is chosen as the evaluation metric, measuring the proportion of correctly classified instances.

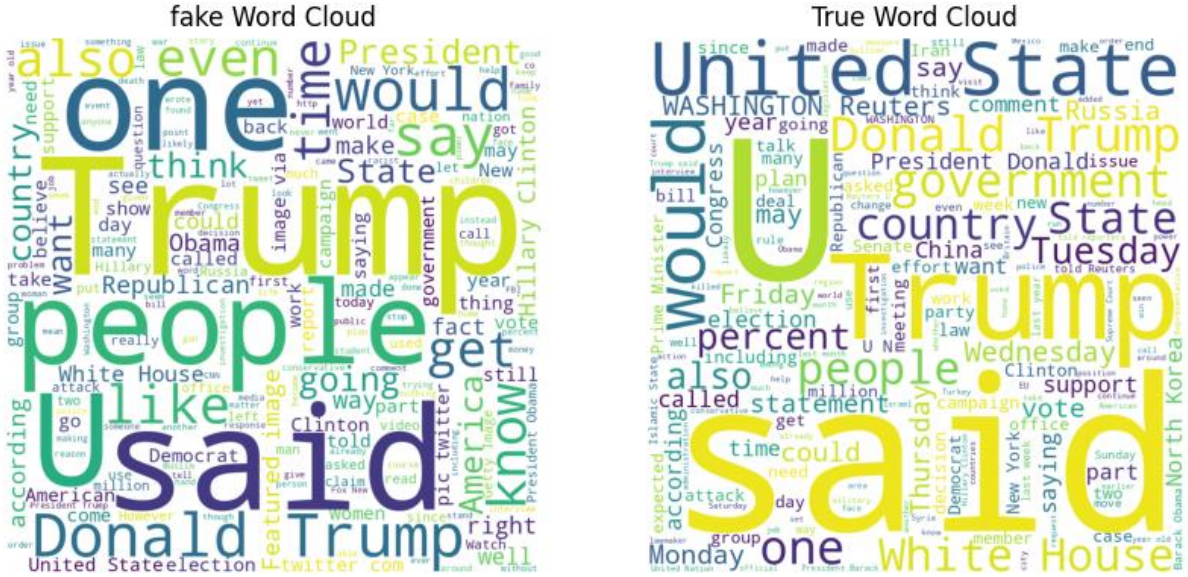
Model Details



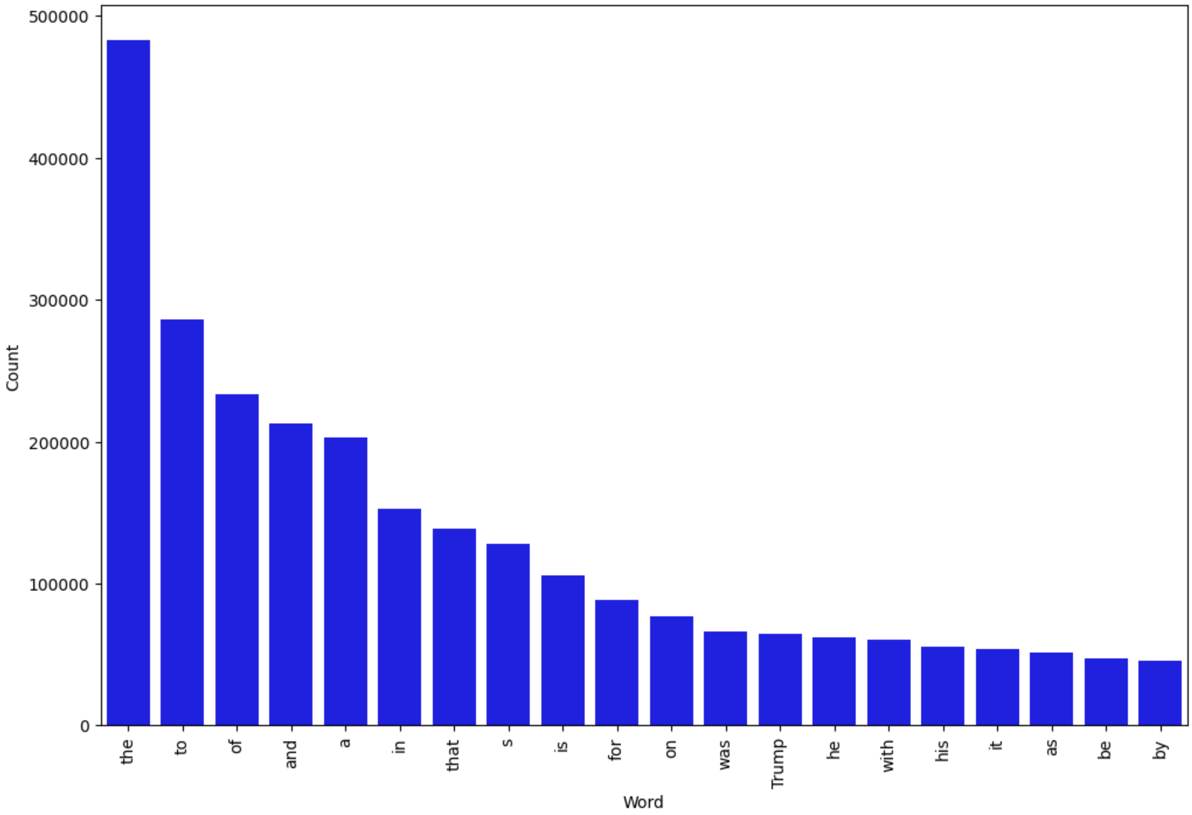
```
Model: "sequential_1"

┌─────────────────────────────────┬──────────────────────────┬────────────────┐
│ Layer (type)                    │ Output Shape             │        Param # │
├─────────────────────────────────┼──────────────────────────┼────────────────┤
│ dense_2 (Dense)                 │ (None, 64)               │      7,393,280 │
├─────────────────────────────────┼──────────────────────────┼────────────────┤
│ dense_3 (Dense)                 │ (None, 1)                │             65 │
└─────────────────────────────────┴──────────────────────────┴────────────────┘

 Total params: 22,180,037 (84.61 MB)

 Trainable params: 7,393,345 (28.20 MB)

 Non-trainable params: 0 (0.00 B)

 Optimizer params: 14,786,692 (56.41 MB)
```



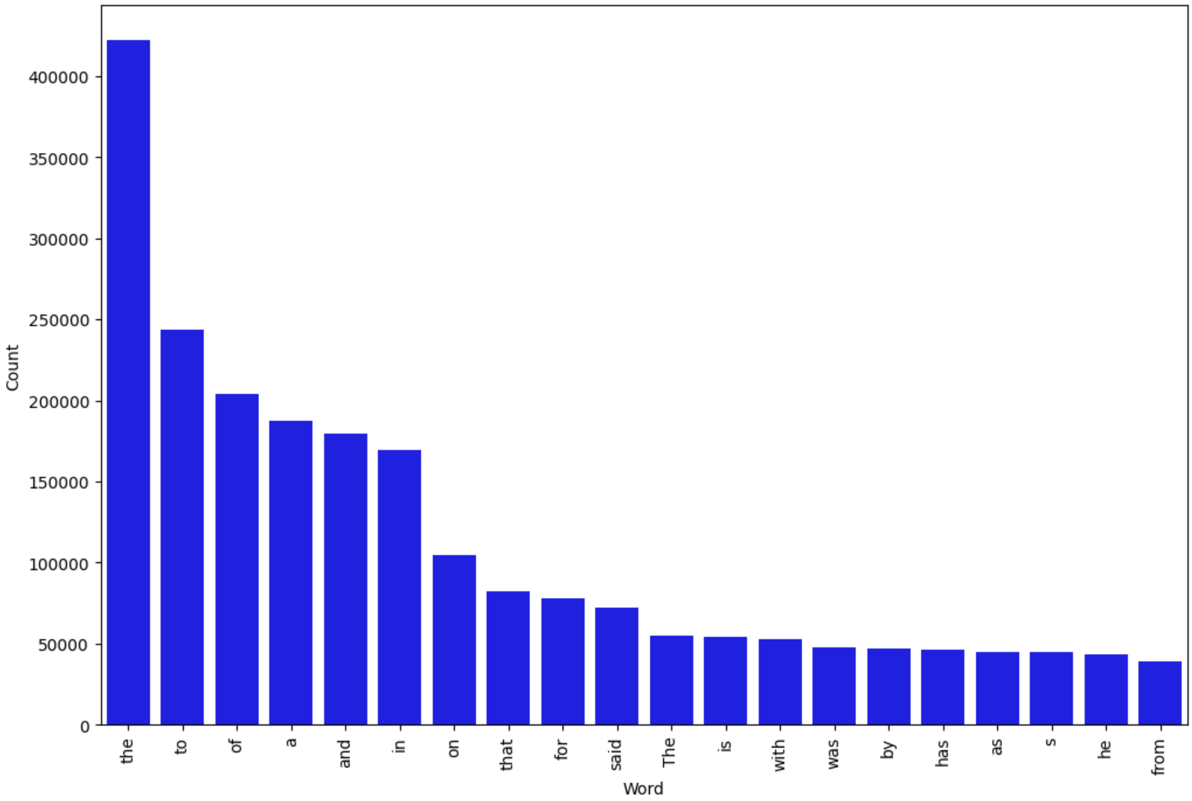




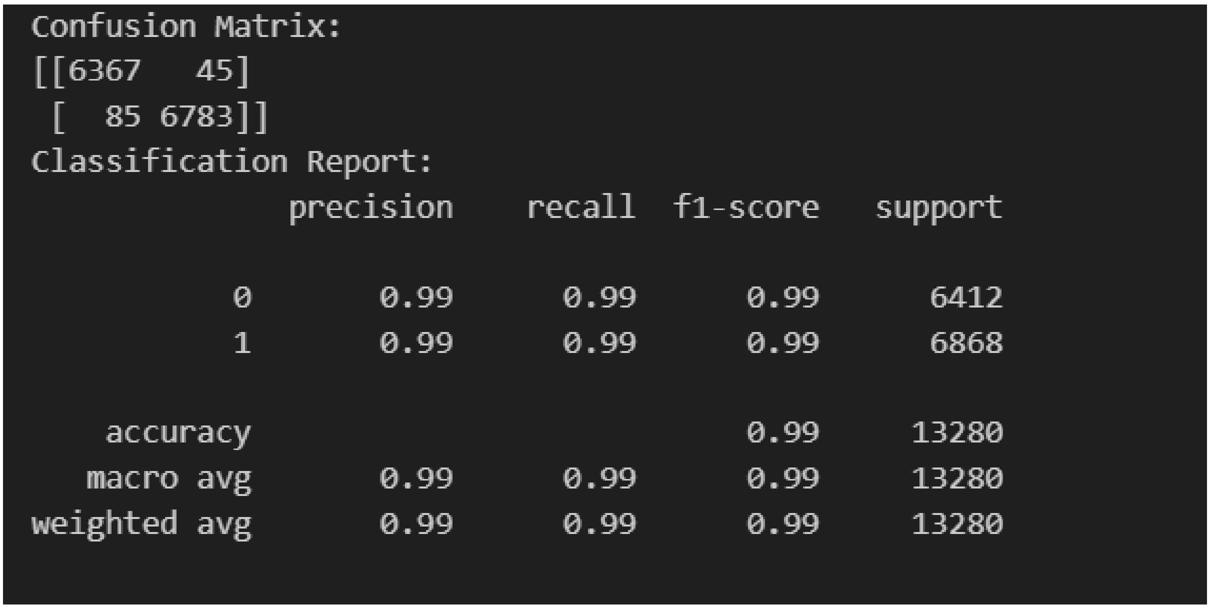
Word Cloud Fake Vs Real



Most frequent words in fake news



Most frequent words in real news



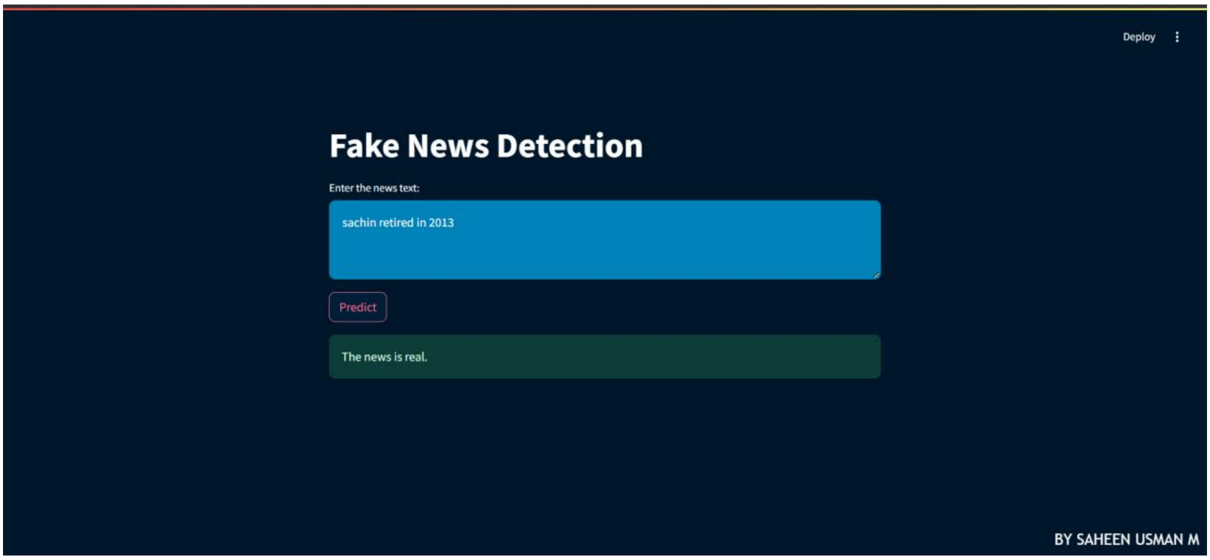
```
Confusion Matrix:
[[6367   45]
 [  85 6783]]
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      6412
           1       0.99      0.99      0.99      6868

    accuracy                           0.99     13280
   macro avg       0.99      0.99      0.99     13280
weighted avg       0.99      0.99      0.99     13280
```

Confusion Matrix & Classification Report

Conclusion

Despite achieving a high accuracy of 99%, the risk of overfitting remains significant. I experimented with various preprocessing techniques and both simpler and more complex models, but the results consistently pointed to potential overfitting. This suggests that while the model performs well on the training data, it may not generalize effectively to unseen data.

The training data, collected from 2016 to 2017, includes 10,145 articles on World News, 11,272 on Politics, and 23,481 labeled as Fake News. Additionally, it encompasses 1,570 Government News articles, 778 on the Middle East, 783 on US News, 4,459 on Left-wing News, and 6,841 on general politics. Given the specific nature and categories of the dataset, the model may not be generalized for all purposes. The varied topics reflect a diverse range of subjects, yet the model's performance is tailored to this particular distribution, limiting its applicability to other contexts.



Fake News Detection

Enter the news text:

sachin retired in 2013

The news is real.

BY SAHEEN USMAN M

A frontend using Streamlit