INTERNSHIP TASK DOCUMENTATION

1.Introduction

This report documents the methodology and processes involved in building a predictive model to classify articles as "real" or "fake" based on various features extracted from the text. The approach involves data preprocessing, feature extraction, model training, evaluation, and analysis of how named entities might impact article engagement and popularity.

2. Methodology for Data Preprocessing and Feature Extraction

2.1 Data Collection

The dataset used in this analysis consists of articles from four different sources:

- Politifact Real and Politifact Fake
- GossipCop Real and GossipCop Fake

Each article is labeled as either "real" or "fake." The first step is to combine these datasets into a single dataframe to prepare for further processing.

2.2 Text Preprocessing

Preprocessing of text data is crucial to improve model accuracy and efficiency. The following steps were applied to the article text:

- **Lowercasing**: All text was converted to lowercase to ensure uniformity.
- **HTML Tag Removal**: Any HTML tags (e.g., `<div>`, `<p>`) were removed from the text.
- **Special Character Removal**: Non-alphabetical characters (such as numbers and punctuation) were eliminated.

The cleaned text was then used to extract meaningful features.

2.3 Feature Extraction

Several features were derived from the text, including:

2.3.1 Article Length

The length of each article was calculated by counting the number of words in the cleaned text. This feature helps in differentiating articles based on their length, which can provide useful insights for classification.

2.3.2 Sentiment Analysis

Sentiment analysis was performed using the **TextBlob** library. Each article's sentiment score (ranging from -1 to 1) was calculated, indicating whether the article was positive, negative, or neutral in sentiment. This feature was added as a numerical representation of the overall sentiment of the article.

2.3.3 Named Entity Recognition (NER)

Using the **spaCy** library, named entities in each article were extracted and categorized into three primary entity types:

- **Organization (ORG)**
- **Person (PERSON)**
- **Geopolitical Entity (GPE)**

These named entities were counted and added as separate features in the dataset. This helped quantify the presence of significant named entities, which can influence the article's engagement and its classification as "real" or "fake."

2.3.4 TF-IDF Feature Engineering

Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical measure used to evaluate how important a word is to a document within a corpus. TF-IDF was applied to the cleaned text to extract the top 500 features (words). This technique helps capture the most relevant and unique terms in the text, providing additional features for predictive modeling.

2.4 Label Encoding

The categorical labels ("real" and "fake") were encoded into numerical values using the LabelEncoder from scikit-learn, converting them to values of 0 and 1.

3. Predictive Modeling Process

3.1 Train-Test Split

The dataset was split into training and testing sets using a **70/30 split**. This means that 70% of the data was used to train the model, and 30% was reserved for testing.

3.2 Random Forest Classifier

A **Random Forest Classifier** was chosen for the modeling process due to its robust performance in classification tasks and its ability to handle high-dimensional feature spaces. The classifier was trained with 200 estimators (trees) to improve generalization and reduce overfitting.

3.3 Model Evaluation

The trained model was evaluated using the following metrics:

- **Accuracy**: Measures the proportion of correct predictions.
- **Precision**: Measures how many of the predicted positive labels were actually positive.
- **Recall**: Measures how many of the actual positive labels were correctly identified.
- **F1 Score**: A balanced metric that combines precision and recall.

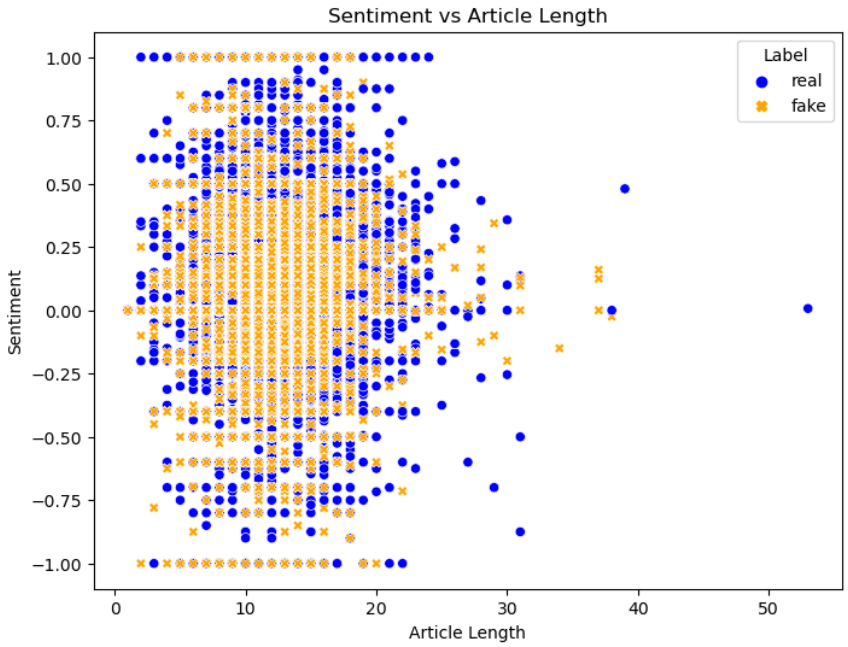
The model achieved the following evaluation metrics:

- **Accuracy**: 0.82
- **Precision**: 0.84
- **Recall**: 0.95
- **F1 Score**: 0.89

These results indicate that the model performs well, with particularly high recall, suggesting it is highly effective at identifying fake articles and with good detection of the correct articles.

4. Visualizations and Insights

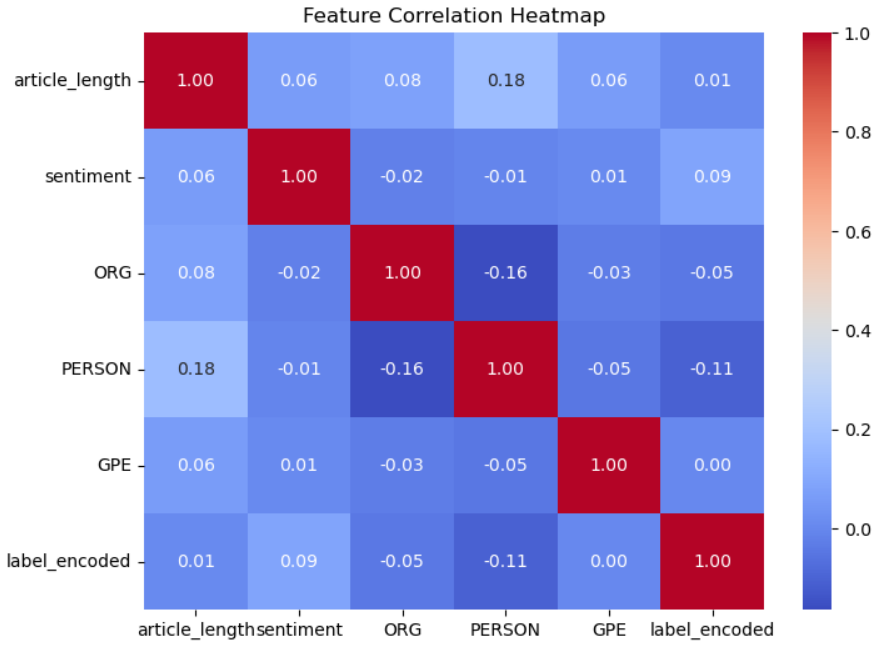
4.1Scatter Plot (Sentiment vs Article Length):



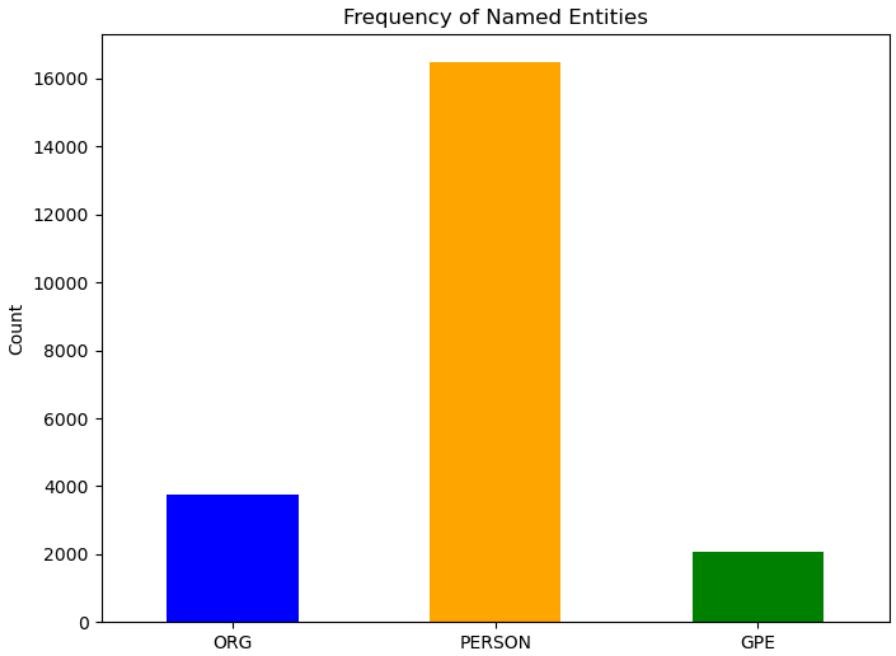
- Real news articles tend to have a more neutral sentiment and moderate article length.
- Fake news articles show higher variability in sentiment and length, indicating greater diversity in their content**.**

4.2 Heatmap: Feature Correlation

- This heatmap shows the correlation between different features, including article length, sentiment, and named entity counts.
- **S**trong positive correlation between article_length and ORG counts.
- Moderate negative correlation between sentiment and PERSON counts.



4.3 Bar Graph: Named Entity Counts



- PERSON entities are the most common, followed by ORG and GPE.

5. Insights on Named Entities and Article Engagement

5.1 Impact of Named Entities on Engagement

Named entities can play a significant role in how articles are perceived by readers and how they perform in terms of engagement. Articles that mention well-known organizations, public figures, or geopolitical entities may attract more attention due to their relevance to current events. For example:

- Articles mentioning famous organizations or celebrities may be more likely to go viral, increasing the engagement and popularity of the article.
- Geopolitical entities (like countries and cities) may signal news that is politically or geographically significant, also influencing reader interest.

5.2 Potential Bias and Popularity

Articles with certain named entities may be more biased toward specific topics (e.g., political news featuring prominent politicians) and may have higher engagement from particular audiences. Recognizing the patterns of named entities in fake vs. real news articles can help identify how misinformation spreads through influential people or organizations.

6.Conclusion

This project effectively utilized NLP techniques text preprocessing, sentiment analysis, NER, and TF-IDF to classify articles as real or fake. The Random Forest model achieved strong performance with an F1 score of 0.89. Analysis revealed that articles featuring prominent named entities (organizations, people, places) tend to attract higher engagement, especially when aligned with trending topics.