Named Entity Recognition (NER) and Feature Engineering

1. Introduction

- **In this report, we analyze a dataset of real and fake news articles to explore the relationship between named entities and article popularity. The primary focus is on how the presence of named entities such as persons, organizations, and geopolitical entities influences the engagement and popularity of news articles.**

2. Methodology

- **Data Loading and Preprocessing:**
  - **We began by loading four datasets: two for real news and two for fake news from sources such as GossipCop and Politifact. Each dataset was labeled accordingly (`1` for real news and `0` for fake news).**
  - **The datasets were then combined, and a new feature, `tweet_count`, was added, which calculates the number of tweet IDs in each article.**
- **Text Preprocessing:**
  - **The text from each article's title was cleaned by removing HTML tags, special characters, numbers, and converting the text to lowercase.**
  - **Stopwords were removed, and the text was tokenized using the `nltk` library to prepare it for further analysis.**
- **Feature Extraction:**
  - **Named Entity Recognition (NER): We used SpaCy's pre-trained model to extract named entities from the cleaned text. Specifically, we extracted `PERSON`, `ORG`, and `GPE` (Geopolitical Entity) counts for each article.**
  - **Article Length: We computed the length of each article in terms of the number of words.**
  - **Sentiment Analysis: Sentiment polarity for each article was computed using the `TextBlob` library, providing a measure of sentiment intensity (positive or negative).**
- **Predictive Modeling:**
  - **A Random Forest Classifier was trained using features such as `person_count`, `org_count`, `gpe_count`, `article_length`, and `sentiment`. The model was used to predict whether an article is real or fake based on these features.**
  - **The model's performance was evaluated using metrics like precision, recall, and F1-score.**

3.Findings:

- **Correlation Analysis:**
  - **We analyzed the correlation between `tweet_count` (a proxy for popularity) and the news type (real vs. fake). The correlation was found to be weak but negative (`-0.051411`), suggesting a slight inverse relationship between tweet count and news type.**
  - **On average, fake news articles had a higher tweet count (132.69) compared to real news articles (74.53), indicating that fake news might engage a larger audience or be shared more widely on social media.**
- **Average Popularity (Tweet Count) for Real vs Fake News:**
  - **The average tweet count for fake news articles is significantly higher than that for real news articles. This indicates that fake news tends to attract more social media engagement.**

4. Visualizations:

- **Bar Chart: Entity Frequency in News Articles:**
  - **This bar chart shows the total frequency of each named entity type (PERSON, ORG, GPE) across both real and fake news articles. We can observe which entities are more prevalent in the articles.**
- **Scatter Plot: Correlation Between Named Entity Counts and Popularity:**
  - **This scatter plot illustrates the relationship between the number of PERSON entities and the tweet count. The points are colored based on whether the news is real or fake, helping to visualize how entity presence correlates with popularity.**
- **Heatmap: Correlation Between Entity Counts and Popularity:**
  - **A heatmap was generated to visualize the correlations between different entity counts (PERSON, ORG, GPE), tweet count, and sentiment. This helps identify any significant patterns in the data.**

5. Insights

- **Impact of Named Entities on Engagement:**
  - **Articles with more named entities, particularly persons and organizations, tend to have higher engagement in terms of tweet counts. This suggests that news articles with recognizable figures or institutions attract more attention.**
  - **The sentiment of an article also plays a role in engagement. Articles with stronger positive or negative sentiments tend to attract more engagement, while neutral articles may have less impact.**

- **Fake News and Popularity:**
    - **Fake news articles tend to have higher tweet counts, which may be due to their more sensationalist or controversial nature. These articles are often shared more frequently on social media, increasing their visibility and engagement.**
- **Model Performance:**
    - **The Random Forest Classifier performed reasonably well in distinguishing between real and fake news based on the features extracted. The model's performance metrics, including precision, recall, and F1-score, were used to evaluate its effectiveness.**

6. Conclusion

- **The analysis reveals a significant relationship between named entities and article engagement. Real and fake news articles show different engagement patterns, with fake news articles attracting more attention.**
- **Named entities such as persons, organizations, and geopolitical entities are crucial in understanding the engagement of news articles, with their presence likely driving more social media interaction.**
- **This analysis can help improve the understanding of news dissemination and engagement patterns, providing insights for fake news detection and content analysis.**