Task 1: Named Entity Recognition (NER) and Feature Engineering

1. Introduction

This study aims to explore the relationship between named entities, sentiment, and article characteristics to distinguish fake news from real news and analyze their impact on engagement and popularity.

The dataset includes articles from **FakeNewsNet**, with labeled real and fake news data from sources such as Politifact and GossipCop.

The primary objectives of this study are to explore the role of named entities and linguistic features in understanding article engagement and popularity. By analyzing the influence of named entities such as PERSON, ORG, and GPE on article performance, the study seeks to determine their impact on readership and popularity metrics. Additionally, it investigates how linguistic features, including sentiment polarity, subjectivity, and title length, correlate with audience engagement. Finally, the study aims to leverage these insights to develop predictive models capable of assessing article popularity, offering a data-driven approach to understanding content performance.

2. Methodology

2.1 Data Preprocessing

- **Data Cleaning**: Titles were cleaned by removing HTML tags, punctuation, numbers, and stopwords to enhance feature extraction.
- **Labeling**: Articles were labeled with is_fake (1 for fake, 0 for real) to facilitate classification.
- **Normalization**: Texts were converted to lowercase and tokenized for uniform analysis.

2.2 Named Entity Recognition (NER)

- **NER Extraction**: SpaCy's pre-trained model identified named entities such as PERSON (individuals), ORG (organizations), and GPE (geopolitical entities).
- **Entity Counts**: Features like num_PERSON, num_ORG, and num_GPE were extracted to capture the frequency of entity mentions in titles.

2.3 Sentiment Analysis

- **Polarity and Subjectivity**: Sentiment scores were calculated using TextBlob:
  - **Polarity** measures the positivity/negativity of the text (range: -1 to 1).
  - **Subjectivity** measures the subjectivity level (range: 0 to 1).

2.4 Feature Engineering

The following features were engineered:

- **Entity Metrics**: num_entities, num_PERSON, num_ORG, num_GPE.
- **Linguistic Features**: Sentiment polarity, subjectivity, and title length.
- **Engagement Metric**: Popularity, derived from the number of tweets, served as the dependent variable in the prediction model.

3. Visualizations and Insights

3.1 Correlation Analysis

- **Heatmap Analysis**: Weak correlations were observed between is_fake and other features:
  - num_PERSON had the highest correlation with is_fake (0.13).
  - num_entities correlated strongly with num_PERSON (0.59), indicating articles mentioning individuals often included other entities.

3.2 Named Entity Frequency

- **Frequency Distributions**: Articles mentioning PERSON entities were more frequent than ORG or GPE, suggesting a preference for human-centered content.

3.3 Polarity and Popularity

- **Sentiment Trends**: Articles with neutral polarity dominated both real and fake news, showing no significant difference in sentiment between the two categories.
- **Popularity vs. Polarity**: Popular articles did not exhibit any distinct polarity trends, highlighting the limited impact of sentiment on engagement.

3.4 Article Length and Popularity

- **Length Distribution**: Article length (measured by word count) showed no significant correlation with popularity.
- **Real vs. Fake**: Both real and fake articles had similar length distributions, indicating that content volume does not determine authenticity.

3.5 Engagement by Entity Types

- **Entity Impact**: Articles with higher numbers of PERSON and ORG entities showed slightly higher engagement, suggesting a potential role of these entities in driving readership.

4. Predictive Modelling

4.1 Logistic Regression

- **Goal**: To classify articles based on engagement (popularity).
- **Features Used**: Sentiment, entity counts, and title length.
- **Performance**: The model demonstrated limited predictive power, highlighting the complex nature of engagement prediction.

4.2 Random Forest Regression

- **Goal**: To predict popularity (numerical engagement score) using the same features.
- **Performance Metrics**:
  - **RMSE**: Measured the model's prediction error.
  - **R-squared (R^2)**: Explained variance, indicating moderate prediction success.

4.3 Model Comparison

- Random Forest outperformed Logistic Regression in predictive accuracy, suggesting non-linear relationships among features.

5. Conclusion

This study underscores the complexities involved in predicting article engagement and classifying fake news using linguistic features and named entities. While features such as named entities like PERSON and ORG show a moderate influence on engagement, they are not sufficient as standalone predictors and require integration with additional contextual and external factors for effective prediction. The findings suggest that more advanced approaches, such as leveraging contextual embeddings (e.g., BERT) and incorporating external variables like user demographics and publication timing, could significantly enhance performance. Overall, the study provides valuable insights but highlights the need for more sophisticated methodologies to achieve robust and accurate predictions.