Named Entity Recognition and Predictive Modeling of Article Popularity

1. Introduction

The objective of this task was to analyze a set of news articles and predict article popularity based on various features, including named entity recognition (NER) and sentiment analysis. We extracted entities from article titles (such as organizations, locations, and people), and engineered additional features like article length and sentiment score to build a predictive model. The goal was to assess how these features affect the engagement and popularity of articles.

2. Methodology

2.1 Data Preprocessing

Data preprocessing is a critical step to ensure the text is clean and structured before feature extraction and model building. The preprocessing steps we performed are:

1. **Cleaning the Text**:
    - Removed HTML tags using regular expressions.
    - Eliminated special characters and numbers (non-alphabetic).
    - Removed extra spaces and normalized text by converting it to lowercase.

The cleaned text from the title column was stored in a new column called cleaned_title.

2. **Tokenization and Stopword Removal**:
    - We tokenized the text (i.e., split the title into individual words) and removed common stopwords (words such as "the", "a", "of", etc.) using NLTK's stopwords list.
3. **Named Entity Recognition (NER)**:
    - We used SpaCy's pre-trained NER model to extract named entities from the article titles. Entities were categorized into three types:
        - **Organizations (ORG)**
        - **Locations (GPE)**
        - **People (PERSON)**

The count of each entity type (i.e., number of organizations, locations, and persons) in each article title was stored as separate features: org_count, location_count, and person_count.

2.2 Feature Extraction

We engineered the following features from the cleaned article titles:

1. **Article Length**:
    - The number of words in the article title (article_length) was used as a feature. This gives an idea of the title's length, which may influence article popularity.
2. **Sentiment Score**:
    - We used the **TextBlob** library to calculate the sentiment score for each article title. Sentiment is expressed as a polarity score, where negative values represent negative sentiment, positive values represent positive sentiment, and zero represents a neutral sentiment.

The sentiment score for each title was stored as the sentiment_score feature.

2.3 Target Variable (Popularity)

For the sake of this task, we assumed that the popularity of an article can be approximated by the number of tweet IDs associated with each article, as indicated in the tweet_ids column. We derived a popularity feature based on the number of tweet IDs associated with each article. This feature acts as a proxy for the engagement and popularity of the article.

3. Predictive Modeling

3.1 Model Selection

For predicting article popularity, we used the **Random Forest Regressor**, which is a powerful ensemble learning method well-suited for regression tasks. This model works well with complex datasets where relationships between features are non-linear.

3.2 Model Training and Evaluation

We split the dataset into training and testing sets (80% for training and 20% for testing) using train_test_split from **scikit-learn**.

The following features were used to train the model:

- org_count (number of organizations)
- location_count (number of locations)
- person_count (number of persons)
- article_length (length of the article title)
- sentiment_score (sentiment score of the article title)

The model was trained on these features and then tested on the hold-out test set. The model's performance was evaluated using **Mean Absolute Error (MAE)**, which measures the average magnitude of errors in predictions.

3.3 Model Results

The **Mean Absolute Error (MAE)** of the Random Forest model was calculated to evaluate its performance. MAE is a useful metric for regression tasks as it gives a clear indication of how far off our predictions are from the actual values. A lower MAE indicates better model performance.

4. Visualizations

We created the following visualizations to help understand the relationships between named entities and article popularity:

4.1 Bar Chart for Entity Frequency

We created a bar chart to visualize the frequency of named entities (organizations, locations, and people) across the entire dataset. This provides insights into the types of entities that are most frequently mentioned in the articles.

- **Insight**: From the chart, we observed that the most frequent entities mentioned were either organizations or locations, indicating that these types of entities are more common in news articles, reflecting their relevance in popular news topics.

4.2 Scatter Plot for Article Length vs Popularity

We used a scatter plot to explore the relationship between the article title length and its popularity (based on tweet counts).

- **Insight**: The plot revealed that there is some variation in article length, but it doesn't show a strong linear relationship with popularity. Longer titles don't necessarily correlate with higher engagement.

4.3 Heatmap for Feature Correlation

We created a heatmap to examine the correlation between different features, including the relationship between named entity counts (organization, location, person) and article popularity.

- **Insight**: The heatmap showed some correlations between sentiment scores and popularity, but entity counts (organization, location, person) did not show strong direct correlations with popularity. However, entities can influence the themes of articles, which indirectly impacts popularity.

5. Insights on Named Entities and Article Engagement

- **Named Entities Impact**: The entities identified in the articles (organizations, locations, and people) provide valuable context for understanding the article's subject matter. While we did not find strong direct correlations between the number of entities and popularity, certain types of articles with prominent people or organizations tend to attract more attention.
- **Entity Frequency**: The frequency of entities suggests that certain organizations or well-known figures are frequently mentioned in the titles, which may drive engagement, though further analysis is needed to understand the direct impact of specific entities.
- **Article Length and Sentiment**: While article length did not show a strong correlation with popularity, sentiment analysis revealed that positive sentiment tends to have a slight positive effect on popularity, suggesting that articles with a more optimistic tone might attract more engagement.

6. Conclusion

This analysis demonstrates how named entity recognition and basic text features (such as article length and sentiment score) can be used to engineer meaningful features for predicting article popularity. While there is no strong linear relationship between named entity counts and popularity, the insights gained from the visualizations suggest that the content of the article (e.g., people, organizations, and locations mentioned) plays a crucial role in shaping engagement.

By using the **Random Forest Regressor**, we were able to model article popularity reasonably well, with **Mean Absolute Error (MAE)** being the primary metric used to assess the model's performance.

Future work could involve refining feature engineering, incorporating more detailed engagement metrics (such as likes, shares, and comments), and using more advanced models for improved prediction.