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

## Methodology for Data Preprocessing and Feature Extraction

### a) Data Cleaning

To guarantee consistency, several preprocessing procedures were applied to the text data. First, regular expressions were used to remove redundant HTML tags and whitespace. Only the words for analysis remained after special characters were removed. Additionally, to preserve consistency and prevent inconsistencies brought on by case sensitivity, all text was transformed to lowercase.

### b) Tokenization and Stopword Removal

Tokenization was used to separate the text into distinct words and so break it down into relevant components. Common stopwords like "the," "is," and "and" were eliminated from the text using the NLTK package. Eliminating these words allowed the analysis to concentrate on the most crucial terms because they didn't provide much meaning.

### c) Named Entity Recognition (NER)

Using spaCy's Named Entity Recognition (NER) tool, named entities were extracted. This made it possible to identify particular textual entities, such as organizations (ORG), geopolitical entities (GPE), and individuals (PERSON). In order to provide useful information for predictive modeling, the number of instances of each entity type was then tallied for each article.

### d) Feature Creation

Along with named entities, two more features were developed: sentiment score (determined by TextBlob) and article length (the number of words in the article). From negative to positive, the emotion score represents the article's overall tone. The engagement metric could not have been constructed without these qualities.

## 2. Predictive Modeling Process and Performance Metrics

### a) Regression Analysis for Engagement Metric Prediction

A Random Forest Regressor was applied to predict article engagement. The goal was to predict a continuous engagement metric based on the named entity counts, sentiment score, and article length. The model's performance was assessed using Mean Absolute Error (MAE), which quantifies the difference between predicted and actual engagement scores. The lower the MAE, the more accurate the model's predictions.

### b) Binary Classification of Engagement

To simplify the engagement prediction, the engagement metric was categorized into two classes: high engagement and low engagement. This was done by comparing each article's engagement score to the median value. Logistic Regression and Random Forest Classifier models were trained to classify the articles into these categories. The performance of these models was evaluated based on accuracy and F1 score, which measure the proportion of correct predictions and the balance between precision and recall, respectively.

**3. Findings from Visualizations and Analysis**

**a) Named Entity Frequency**

A bar chart was created to visualize the frequency of different named entities (ORG, GPE, and PERSON) across all articles. The results indicated that geopolitical entities (GPE) were the most frequently mentioned, followed by organizations (ORG) and persons (PERSON). This highlighted the importance of location and organizational references in the dataset.

**b) Sentiment vs. Engagement**

A scatterplot examined the relationship between sentiment score and engagement metric. The plot showed a positive correlation, suggesting that articles with a more positive sentiment tend to have higher engagement scores. This indicates that the tone of the article is an important factor in reader engagement.

**c) Correlation Analysis**

A heatmap was generated to visualize the correlations between different features, including named entity counts and the engagement metric. The heatmap revealed strong correlations, especially between the number of ORG and GPE entities and engagement. This suggests that articles mentioning these entities are likely to attract more engagement from readers.

**4. Insights on Named Entities and Article Engagement**

**a) The Role of Named Entities**

The analysis revealed that articles mentioning specific named entities, such as organizations, people, or places, tend to receive higher engagement. This is likely due to the relevance and interest that such entities generate among readers. Articles that focus on current events, notable individuals, or well-known organizations are more likely to engage a larger audience.

**b) The Impact of Sentiment**

Sentiment research revealed that articles with a positive tone were linked to higher levels of interaction. This implies that information having an upbeat tone has a higher chance of interacting with readers. To increase reader engagement, content producers could find it helpful to concentrate on uplifting tales.

**c) Practical Implications for Content Strategy**

Based on the findings, publishers and content creators should consider the inclusion of named entities, particularly organizations and locations, in their articles. Additionally, maintaining a positive sentiment can further increase engagement. Tailoring content to include these elements could help drive more traffic and interaction on their platforms.