

Fake News Analysis Report

1. Methodology

1.1 Data Preprocessing

The dataset used for this analysis was loaded from a CSV file containing fake news articles. The preprocessing steps included:

1. **Data Loading:** The dataset was loaded using pandas. A function was defined to handle potential file loading errors gracefully.
2. **Title Extraction:** Titles of articles were extracted for performing Named Entity Recognition (NER).
3. **Tweet IDs:** These were used to calculate popularity by counting the number of tweet IDs associated with each article. Missing or invalid entries were handled using pandas' apply method.

1.2 Feature Extraction

Named Entity Recognition (NER)

The spaCy library's pre-trained model (en_core_web_sm) was used for NER. Key steps included:

1. **Entity Identification:** Entities were extracted from article titles.
2. **Entity Counting:** The total number of entities per title was calculated.

Popularity Calculation

Popularity was determined by counting the number of tweet IDs associated with each article. This provided an estimate of article engagement.

Additional Features

- **Entity Count:** Sum of all named entities in a title.
- **Is Fake:** A boolean column was added to indicate that all articles in this dataset are fake (as per dataset characteristics).

2. Predictive Modeling Process

2.1 Objective

The objective was to analyze the relationship between named entities and article popularity to gain insights into how named entities affect engagement.

2.2 Visualization

Scatter plots were used to visualize the relationship between entity count and popularity:

- **X-axis:** Entity count (sum of named entities in the title).
- **Y-axis:** Popularity (number of tweet IDs).
- **Hue:** News type (fake, as all entries in the dataset are fake).

This provided a clear picture of whether a higher number of named entities correlates with increased engagement.

3. Performance Metrics

No predictive modeling or classification was performed as this analysis was primarily exploratory. Future extensions could involve:

- Building a classification model to predict article engagement based on named entities and other features.
 - Evaluating models using metrics such as accuracy, precision, recall, and F1-score.
-

4. Insights

Named Entities and Engagement

- Articles with a higher count of named entities in their titles tended to show increased popularity, as measured by tweet counts.
- The presence of named entities may attract more attention and shares, suggesting their importance in crafting engaging headlines.

Implications for Fake News Detection

- Named entities could be a key feature in distinguishing fake from genuine news when combined with other contextual features.

