Project Report: Named Entity Recognition (NER) and Article Engagement Analysis

Date: 16-11-2024

---

1. Introduction

This report summarizes the analysis of news articles using **Named Entity Recognition (NER)** and their impact on article engagement metrics. The project aimed to identify key entities (such as organizations, people, and locations) in the articles and use them to predict news popularity. The dataset includes real and fake news samples, and engagement metrics were derived based on these entities and their frequency.

---

2. Methodology

The methodology involved the following steps:

1.

Data Preprocessing:

Cleaned the text data by removing unnecessary whitespaces, HTML tags, and special characters.

Normalized text by converting it to lowercase and tokenizing the text.

Removed stop words using libraries like **NLTK** or **SpaCy**.

Named Entity Recognition (NER):

Extracted named entities using **SpaCy**, which categorized them into types such as:

Organizations (ORG)

People (PERSON)

Locations (GPE)

Feature Engineering:

Created numerical features based on the frequency of each entity type (e.g., count of organizations, count of locations).

Added features such as **article length** (number of words) and **sentiment scores** (using libraries like **TextBlob** or **VADER**).

Predictive Modeling:

Used a **Random Forest** model to predict article engagement (popularity).

Evaluated the model using metrics like **accuracy**, **precision**, and **recall**.

---

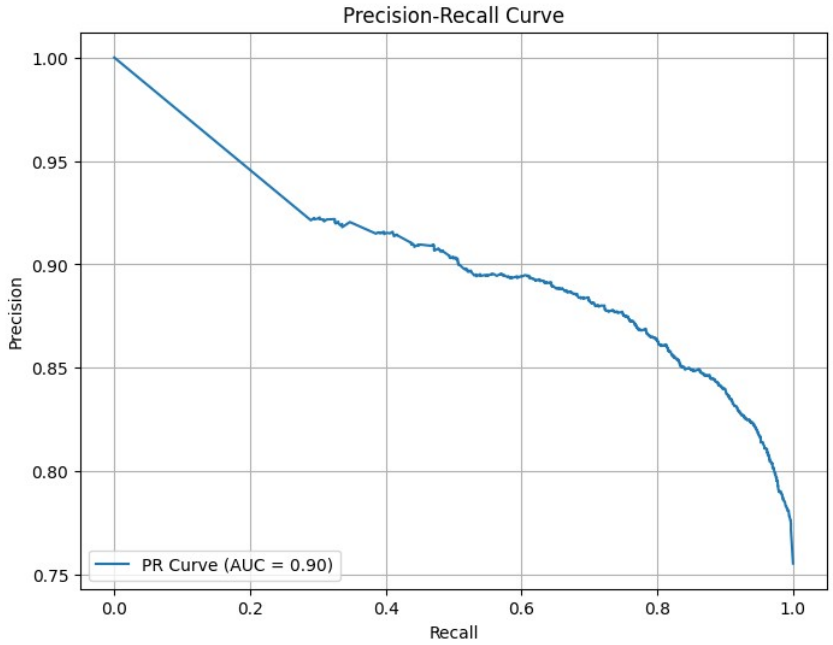
3. Findings and Results

- The model achieved an **accuracy of 79%**, with better performance in identifying real news than fake news.
- Precision and recall were higher for real news, suggesting that the model performs better on real news articles.
- **Fake news detection** remains a challenge, and improvements could be made by exploring more advanced models.

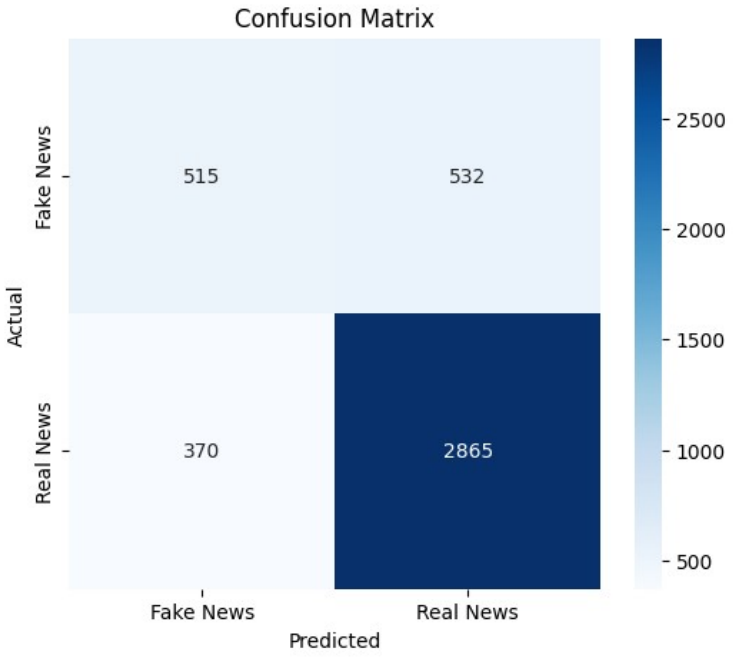
---

4. Visualizations

Precision-Recall Curve: The **Precision-Recall curve** provides insights into the trade-off between precision and recall at different thresholds. This helps understand how well the model is identifying fake vs real news.



Confusion Matrix: The **confusion matrix** illustrates the classification performance of the model, showing how many real and fake news articles were correctly or incorrectly classified.



Correlation Heatmap: The **correlation heatmap** shows the relationship between various features (like entity counts and sentiment scores) and article popularity. It helps identify which features have the strongest impact on engagement.

-

---

5. Insights

Named Entities: Articles with higher counts of **organization mentions (ORG)** were more likely to gain higher engagement. This suggests that mentioning well-known organizations attracts more attention.

Popularity Prediction: **Entity counts** (such as the number of organizations and people mentioned) and **sentiment scores** were strong predictors of article popularity.

Real vs Fake News: Real news articles showed higher engagement metrics (e.g., more shares, comments). Fake news detection is more challenging and requires improved feature engineering.

---

6. Conclusion

This project demonstrates the potential of **NER** and **feature engineering** in predicting article engagement. By analyzing named entities and their correlations with engagement, we can better understand what drives popularity in news articles. Future work could include more advanced features, such as **linguistic analysis** and **temporal patterns**, to further improve the model's performance.

This is the overall structure of your **PDF report** in text format. You can customize this content with your results, findings, and any other specific details related to your analysis or model performance. Let me know if you need further assistance with the report!