A

project Report

On

Predicting News Article Popularity Using NER

Submitted By:

Nitin Gangwar

Introduction :

In the ever-expanding world of online news, predicting the popularity of articles is a critical task for content creators, marketers, and platforms alike. Popularity often correlates with reader engagement, such as shares, likes, and comments, which can ultimately determine the success and reach of an article. This analysis aims to leverage **Named Entity Recognition (NER)** as a key tool in predicting the popularity of news articles. By identifying and extracting key entities—such as organizations, locations, and people—from article titles, we can uncover patterns that drive engagement. Using a combination of four datasets—**Politifact Real**, **Politifact Fake**, **Gossipcop Real**, and **Gossipcop Fake**—the analysis also considers the impact of fake vs. real news articles on popularity.

The datasets are classified into two main categories: **real news** and **fake news**, allowing us to analyze how these categories influence engagement and article popularity. By applying NER to the article titles, we extract essential features based on the frequency of named entities, such as the number of mentions of organizations (ORG), locations (GPE), and people (PERSON). Additional features like article length and tweet count further augment the dataset. Using these features, a **predictive model** is built to forecast article popularity, followed by performance evaluation through various metrics.

Objective :

The primary objective of this analysis is to develop a **predictive model** that can estimate the **popularity** of news articles based on several features, including named entities. Specifically, the analysis aims to:

1. **Apply Named Entity Recognition (NER)**: Use state-of-the-art models to identify and categorize entities such as organizations, locations, and people within news articles.

2. **Feature Engineering**: Create additional features like tweet counts, article length, and sentiment, along with named entity counts, to enhance the predictive power of the model.

3. **Predict Popularity**: Use the engineered features to train a predictive model capable of classifying articles as either popular or not based on social media engagement.

4. **Evaluate Model Performance**: Assess the model's performance using key metrics such as accuracy, precision, recall, and F1-score.

5. **Analyze Fake vs. Real News**: Explore how the type of news (real vs. fake) influences the engagement and popularity of articles.

2. Methodology :

2.1. Data Collection

We start by combining four datasets that contain information about news articles:

- **Politifact Real News**
- **Politifact Fake News**
- **Gossipcop Real News**
- **Gossipcop Fake News**

Each dataset contains the following columns:

- **title**: The article's headline.
- **url**: The URL of the article.
- **tweet_ids**: IDs representing tweets about the article.
- **label**: Indicates if the article is "real" (1) or "fake" (0).

The datasets are loaded, concatenated, and labeled to indicate whether the article is real or fake.

2.2. Data Preprocessing

2.2.1. Text Cleaning

The text of each article title is cleaned by:

1. **Removing HTML Tags**: Any HTML tags are removed using BeautifulSoup to ensure only the content remains.
2. **Special Character Removal**: Non-alphanumeric characters are stripped using regular expressions.
3. **Lowercasing**: All text is converted to lowercase to standardize it and reduce noise.

2.2.2. Feature Extraction

We extract key features from the cleaned titles using the following methods:

1. **Named Entity Recognition (NER)**:
   - We use the Hugging Face transformers library with a pre-trained BERT model (dbmdz/bert-large-cased-finetuned-conll03-english) to identify named entities in the article titles.
   - **Entity Types**: Organizations (ORG), Geopolitical entities (GPE), and People (PERSON) are the primary entity types extracted.
   - The number of entities of each type (organization, location, person) is counted for each article and used as features.
2. **Popularity Features**:

- The **tweet_count** is calculated from the tweet_ids column. If the article has multiple tweet IDs, each tweet ID is treated as a unique tweet.
- A threshold is set for popularity based on the **median tweet count**. Articles with tweet counts greater than the median are labeled as **popular**.

3. **Fake/Real Classification**:
   - The articles are categorized as either "fake" or "real" based on the dataset (Politifact and Gossipcop).

3. Feature Engineering :

After preprocessing, we derive the following features:

1. **Entity Counts**:
   - **org_count**: Number of organizations mentioned.
   - **gpe_count**: Number of geopolitical entities (locations) mentioned.
   - **person_count**: Number of people mentioned.
2. **Popularity Label**:
   - Articles are classified as either **popular** or **not popular** based on the tweet count exceeding the median.
3. **Fake News Label**:
   - Articles are classified as **real** (1) or **fake** (0) based on their source.

These features will be used to train a predictive model.

4. Predictive Modeling :

4.1. Model Selection

We used the **Random Forest Classifier**, a versatile machine learning model, for predicting article popularity. The Random Forest model was selected due to its ability to handle both numerical and categorical data and its robustness to overfitting.

4.2. Data Splitting

The dataset was split into training and testing sets:

- **Training Set**: 80% of the data.
- **Testing Set**: 20% of the data.

4.3. Model Training

The model was trained on the features:

- **org_count**
- **gpe_count**
- **person_count**
- **label** (real/fake news)

4.4. Model Evaluation

The model's performance was evaluated using the following metrics:

- **Accuracy**: Measures the percentage of correct predictions.
- **F1-Score**: Harmonic mean of precision and recall, useful for imbalanced datasets.
- **Confusion Matrix**: Provides a detailed breakdown of true positives, true negatives, false positives, and false negatives.

4.5. Results

The model performed well with a balanced F1-score, and the confusion matrix revealed that the model correctly identified popular articles in most cases, although there was some misclassification due to the complexity of the features.

5. Insights and Observations :

5.1. Impact of Named Entities on Article Popularity

The analysis revealed the following insights about the relationship between named entities and article popularity:

1. **Organizations (ORG)**: Articles mentioning high-profile organizations tend to have higher popularity. This is likely because readers are more interested in articles about well-known companies, institutions, or corporations.
2. **Geopolitical Entities (GPE)**: Articles that mention well-known locations or cities also appear to be more popular, as locations related to current events (e.g., elections, international crises) drive greater engagement.
3. **People (PERSON)**: Articles mentioning public figures or celebrities tend to attract more attention and shares, contributing to higher popularity.

5.2. Fake vs. Real News

We observed that:

1. **Real News**: Articles labeled as real news tend to have higher tweet counts and engagement. Real news articles about current events or political topics, particularly those from trusted sources, receive more social media attention.
2. **Fake News**: Fake news articles often have a lower tweet count, which could be due to the fact that false information might be shared less or get flagged more quickly.

5.3. Named Entities and Fake News

Named entities can serve as indicators for detecting fake news:

- Fake news articles often contain references to fake organizations or misrepresented locations.
- Real news tends to mention verified organizations, well-known public figures, and locations that align with actual events.

6. Conclusion :

In this analysis, we successfully predicted the popularity of news articles using named entity recognition (NER) and feature engineering. The use of Hugging Face’s **BERT-based NER model** allowed us to extract meaningful entities that correlate with article engagement. Additionally, the model showed how certain named entities (organizations, locations, people) affect the popularity of articles.

By training a Random Forest Classifier and incorporating NER-based features, we gained valuable insights into the impact of named entities on article engagement, providing a foundation for improving the accuracy of popularity prediction in news articles.

7. Future Work :

1. **Model Improvement**: Further tuning of the model (e.g., hyperparameter optimization) could improve performance.
2. **Use of More Features**: Incorporating additional features, such as sentiment analysis, could further enhance the model's predictive capabilities.
3. **Expanding NER**: Including more entity types (e.g., events, products) could refine the model's predictions.
4. **Fake News Detection**: A more sophisticated fake news detection model could be developed using deeper features from the content or source credibility.