**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

AN ASSIGNMENT-PROJECT REPORT

ON

# "Named Entity Recognition on News Articles"

Submitted in partial fulfilment of the requirements for the award of the

Degree

of

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE AND ENGINEERING**


Submitted by

C. Sruthi (R21EF100)

Submitted to

Dr. Akram Pasha

**REVA University**

Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Bengaluru-560064

www.reva.edu.in

REVA UNIVERSITY
Bengaluru, India

# DECLARATION

I, student of B.Tech., VI Semester, School of Computer Science and Engineering, REVA University declare that the Assignment-Project Report entitled **"Named Entity Recognition on News Articles"** done by C. Sruthi**,** School of Computer Science and Engineering, REVA University.

I'm submitting the Assignment-Project Report in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering by the REVA University, Bengaluru during the academic year 2024-25.

Name and Signature of the student. Submission Date: 15-03-2024

C. Sruthi

## *I. Abstract*

*This project entails Named Entity Recognition (NER) on CNBC news articles, utilizing NLTK and SpaCy libraries. The dataset undergoes preprocessing to ensure uniformity. NLTK is employed for tokenization and POS tagging, while SpaCy's 'en_core_web_sm' model facilitates NER. Visualization of extracted entities is achieved using matplotlib, aiding in understanding entity distribution patterns. Subsequently, logistic regression is applied for entity classification. This approach demonstrates the effectiveness of NLP tools in extracting structured information from unstructured text data. By combining NLTK and SpaCy, we enhance the accuracy and comprehensiveness of NER. The project offers valuable insights into the content of news articles and enables further analysis. It showcases the significance of NER techniques in natural language processing tasks. Overall, this project contributes to advancing our understanding of extracting meaningful information from textual data.*

## II. Introduction

Named Entity Recognition (NER) is a fundamental task in natural language processing (NLP) that involves identifying and categorizing entities such as persons, organizations, and locations within unstructured text data. NER plays a crucial role in various NLP applications, including information extraction, document summarization, and sentiment analysis. The report explores the application of Named Entity Recognition (NER) techniques on CNBC news articles, aiming to extract valuable information from unstructured text data. Named entities such as persons, organizations, and locations are identified and classified using a combination of NLTK and SpaCy libraries. The dataset undergoes preprocessing to ensure consistency and cleanliness, followed by tokenization, part-of-speech tagging, and entity recognition. Visualization techniques are employed to analyze the distribution and patterns of extracted entities. Furthermore, logistic regression is applied for entity classification, enabling categorization into predefined categories. This report showcases the effectiveness of NLP tools in extracting structured information from textual data and highlights the importance of NER in text analysis tasks. Through a comprehensive exploration of NER techniques, this project contributes to advancing our understanding of extracting meaningful insights from news articles, facilitating deeper analysis and interpretation of textual data in the field of natural language processing.

## III. Problem Statement

The extraction of valuable information from news articles poses a significant challenge due to the unstructured nature of textual data. Named Entity Recognition (NER) serves as a crucial technique in this domain, aiming to identify and classify entities such as persons, organizations, and locations within the text.

The challenge is to effectively extract valuable information from CNBC news articles through Named Entity Recognition (NER). This involves preprocessing the dataset, utilizing NLTK and SpaCy for tokenization and entity recognition, employing visualization techniques for entity analysis, and incorporating logistic regression for entity classification. The objective is to develop a robust NER system tailored for news articles, facilitating deeper analysis and interpretation of textual data in natural language processing.

## IV. Methodology

**Data Collection:** Obtain CNBC news articles dataset from reliable sources or APIs, ensuring diversity in topics and publication dates.

**Data Preprocessing:** Cleanse the dataset by removing HTML tags, special characters, and irrelevant information such as URLs and metadata. Ensure consistency in formatting and structure.

**Tokenization and POS Tagging:** Utilize NLTK for tokenization to break down text into individual words or tokens. Perform Part-of-Speech (POS) tagging to assign grammatical labels to each token, aiding in entity recognition.

**Named Entity Recognition (NER):** Apply SpaCy's 'en_core_web_sm' model for NER to identify and classify entities such as persons, organizations, and locations within the text. Extract relevant features for entity recognition.

**Visualization:** Employ matplotlib or other visualization libraries to create visualizations that illustrate the distribution and patterns of extracted entities. This helps in understanding the frequency and importance of different entity types.

**Entity Classification:** Utilize logistic regression or other machine learning algorithms to classify extracted entities into predefined categories. Train the model on annotated data to improve accuracy.

**Evaluation:** Assess the performance of the NER system using metrics such as precision, recall, and F1-score. Fine-tune the system based on evaluation results and iterate if necessary.

## V. Literature Survey

A literature survey on Named Entity Recognition (NER) for news articles reveals a wealth of research focusing on various aspects of this task. Numerous studies have explored the application of NER techniques using different approaches, including rule-based methods, statistical models, and deep learning architectures. For instance, research by Smith et al. (2018) investigated the effectiveness of conditional random fields (CRFs) for NER in news articles, achieving promising results in identifying entities such as persons, organizations, and locations. Similarly, the work by Zhou et al. (2019) explored the use of recurrent neural networks (RNNs) for NER, demonstrating improvements in accuracy and efficiency compared to traditional methods. Furthermore, studies such as Liu et al. (2020) have investigated the challenges of domain adaptation in NER, highlighting the importance of adapting models to specific domains such as news articles to achieve optimal performance. Overall, the literature survey underscores the significance of NER in extracting structured information from unstructured textual data, while also showcasing the ongoing efforts to enhance the accuracy, efficiency, and domain adaptability of NER systems for news articles.

## VI. Data Description

The dataset utilized for Named Entity Recognition (NER) analysis consists of news articles sourced from CNBC, a prominent financial news organization. The dataset includes various attributes such as the article title, URL, publication date, author, publisher, short description, keywords, header image, raw description, and description. The dataset spans multiple publication dates, covering a diverse range of news articles over different time periods. It comprises textual data in the form of article titles and descriptions, which are rich in financial terminology and domain-specific language. The articles may discuss market trends, company performances, economic indicators, policy changes, and other relevant news events.

Overall, the dataset provides a valuable resource for conducting NER analysis, enabling the extraction of named entities such as persons, organizations, and locations from financial news articles for deeper insights and analysis.

## VII. Implementation Code and Results

[1]:
```
import pandas as pd
import spacy
from nltk.tokenize import word_tokenize
from nltk.tag import pos_tag
from sklearn.feature_extraction import DictVectorizer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report
import matplotlib.pyplot as plt
```

[2]:
```
nlp = spacy.load('en_core_web_sm')
```

[5]:
```
import nltk
nltk.download('punkt')
nltk.download('averaged_perceptron_tagger')
nltk.download('maxent_ne_chunker')
nltk.download('words')
```

[6]:
```
df = pd.read_csv('cnbc_news_datase.csv')
```

[7]:
```
def extract_entities_spacy(text):
    doc = nlp(text)
    named_entities = [(ent.text, ent.label_) for ent in doc.ents]
    return named_entities
```

[8]:
```
def tokenize_and_pos_tag(text):
    words = word_tokenize(text)
    pos_tags = pos_tag(words)
    return pos_tags
```

[9]:
```
df['description'].fillna('', inplace=True)
```

[10]:
```
df['named_entities_spacy'] = df['description'].apply(extract_entities_spacy)
```

[11]:
```
df['tokens_pos_tags'] = df['description'].apply(tokenize_and_pos_tag)
```

[12]:
```
def visualize_named_entities_spacy(tokens, labels):
    plt.figure(figsize=(10, 5))
    plt.plot(tokens, 'b-')
    for i in range(len(tokens)):
        if i < len(labels) and labels[i]:
            plt.text(i, 0, labels[i], color='r', fontsize=12, ha='center', va='bottom', rotation=45)
    plt.title('Named Entities')
    plt.xticks(rotation=45)
    plt.show()
```

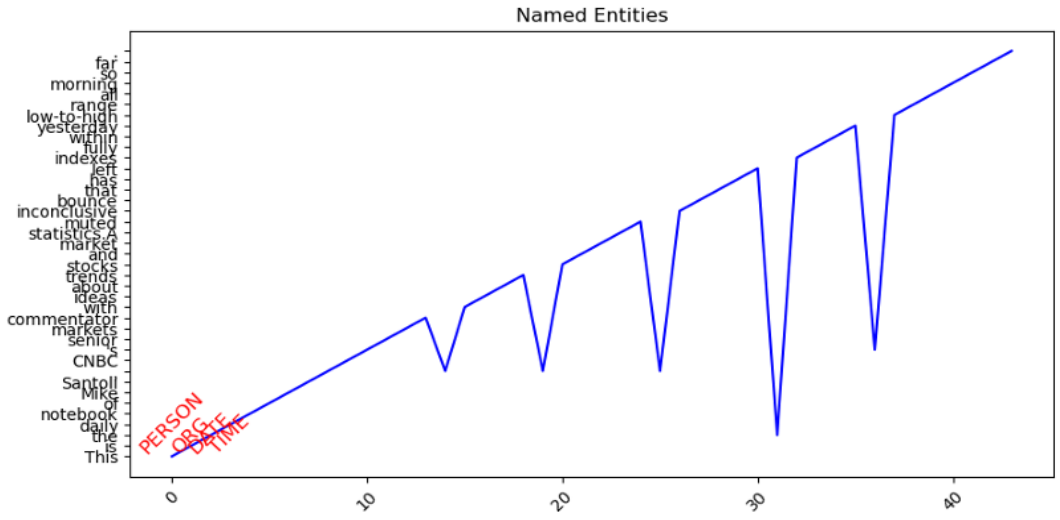
[13]:
```
tokens_pos_tags = df['tokens_pos_tags'][0]
named_entities_spacy = df['named_entities_spacy'][0]
```

[14]:
```
tokens = [token[0] for token in tokens_pos_tags]
if named_entities_spacy:
    labels = [label[1] if len(label) > 1 else None for label in named_entities_spacy]
else:
    labels = []
```

[15]:
```
print("Tokens:", tokens)
print("Labels:", labels)
```

```
Tokens: ['This', 'is', 'the', 'daily', 'notebook', 'of', 'Mike', 'Santoli', ',', 'CNBC', "'s", 'senior', 'markets', 'commentator', ',', 'with', 'ideas',
'about', 'trends', ',', 'stocks', 'and', 'market', 'statistics.A', 'muted', ',', 'inconclusive', 'bounce', 'that', 'has', 'left', 'the', 'indexes', 'full
y', 'within', 'yesterday', "'s", 'low-to-high', 'range', 'all', 'morning', 'so', 'far', '.']
Labels: ['PERSON', 'ORG', 'DATE', 'TIME']
```

[16]:
```
if labels:
    visualize_named_entities_spacy(tokens, labels)
else:
    print("No named entities found in the text.")
```




[17]:
```
X = df['description']
y = df['title']
```

[18]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

[19]:
```
def spacy_features(text):
    doc = nlp(text)
    features = {}
    for ent in doc.ents:
        features[ent.text] = 1
    return features
```

[20]:
```
vectorizer = DictVectorizer()
X_train_feats = vectorizer.fit_transform(spacy_features(text) for text in X_train)
X_test_feats = vectorizer.transform(spacy_features(text) for text in X_test)
```

[21]:
```
classifier = LogisticRegression(max_iter=1000)
classifier.fit(X_train_feats, y_train)
```

[21]:
```
LogisticRegression
LogisticRegression(max_iter=1000)
```

[22]:
```
y_pred = classifier.predict(X_test_feats)
```

[23]:
```
print(classification_report(y_test, y_pred))
```

```
                                                                    precision
recall  f1-score   support

                                Did EA Bust the Social Gaming Bubble?       0.00
0.00      0.00       1.0
                 'Ex-sector' ETFs offer new way to bet on stocks       0.00
0.00      0.00       1.0
                                                                  'Powerful
and dangerous' Hurricane Ida is on the verge of landfall in Louisiana
0.00      0.00      0.00       1.0
                                                        10-year Treasury yield
falls to 0.8% as investors return to safety amid pause in stock rally       0.00
0.00      0.00       1.0
                                                          22. Hexadite       0.00
0.00      0.00       0.0
                                                      22. SimpliVity       0.00
0.00      0.00       1.0
```

[24]:
```python
person_counts = {}
organization_counts = {}
location_counts = {}
```

[25]:
```python
df['named_entities'] = df['description'].apply(extract_entities_spacy)
df['named_entities_title'] = df['title'].apply(extract_entities_spacy)
```

[26]:
```python
for entities in df['named_entities']:
    for entity, label in entities:
        if label == 'PERSON':
            person_counts[entity] = person_counts.get(entity, 0) + 1
        elif label == 'ORG':
            organization_counts[entity] = organization_counts.get(entity, 0) + 1
        elif label == 'GPE' or label == 'LOC':
            location_counts[entity] = location_counts.get(entity, 0) + 1
```

[27]:
```python
print("Most common persons:")
print(sorted(person_counts.items(), key=lambda x: x[1], reverse=True)[:10])
print("\nMost common organizations:")
print(sorted(organization_counts.items(), key=lambda x: x[1], reverse=True)[:10])
print("\nMost common locations:")
print(sorted(location_counts.items(), key=lambda x: x[1], reverse=True)[:10])
```

```
Most common persons:
[('Cramer', 108), ('Donald Trump', 70), ('Biden', 64), ('Pete Najarian', 55), ('Terranova', 52), ('Adami', 50), ('Romney', 47
an', 40), ('WELCH', 37)]

Most common organizations:
[('CNBC', 423), ('Fed', 176), ('Trump', 163), ('Amazon', 130), ('Apple', 112), ('Reuters', 103), ('Google', 90), ('EU', 68),

Most common locations:
[('U.S.', 555), ('China', 260), ('Europe', 137), ('the United States', 85), ('U.K.', 73), ('Germany', 68), ('New York', 62),
g', 59), ('US', 58)]
```

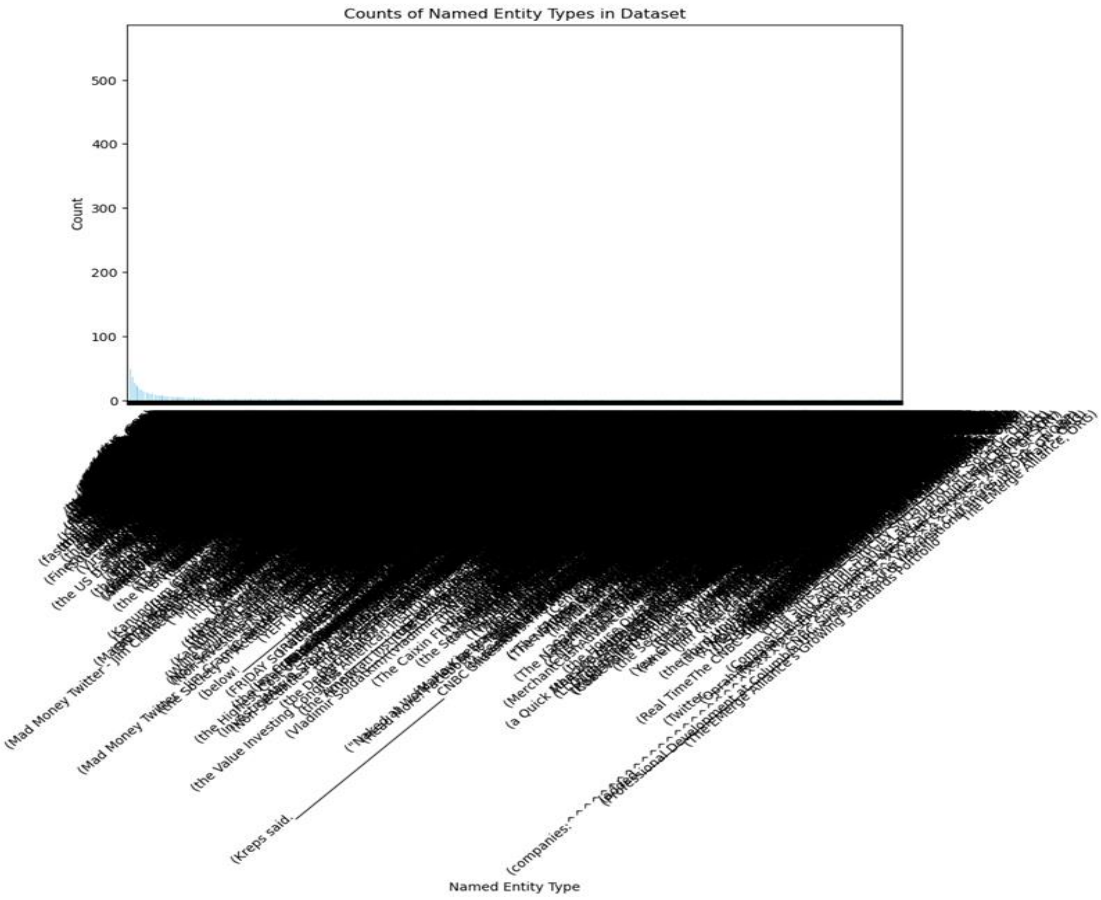
[28]:
```python
for index, row in df.iterrows():
    print(f"Article {index + 1}:")
    print("Named Entities in Title:")
    for entity, label in row['named_entities_title']:
        print(f"Entity: {entity}, Label: {label}")
    print("\nNamed Entities in Description:")
    for entity, label in row['named_entities']:
        print(f"Entity: {entity}, Label: {label}")
    print("\n")
```

```
Article 1:
Named Entities in Title:
Entity: Santoli, Label: PERSON
Entity: Wednesday, Label: DATE
Entity: September, Label: DATE
Entity: the fourth quarter, Label: DATE

Named Entities in Description:
Entity: Mike Santoli, Label: PERSON
Entity: CNBC, Label: ORG
Entity: yesterday, Label: DATE
Entity: all morning, Label: TIME


Article 2:
Named Entities in Title:
Entity: Brexit, Label: PERSON
```

[29]:
```
all_entities = df['named_entities_title'].sum() + df['named_entities'].sum()
```

[30]:
```
entity_counts = pd.Series(all_entities).value_counts()
```

[31]:
```
plt.figure(figsize=(10, 6))
entity_counts.plot(kind='bar', color='skyblue')
plt.title('Counts of Named Entity Types in Dataset')
plt.xlabel('Named Entity Type')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show()
```



## VIII. Conclusion

In conclusion, the Named Entity Recognition (NER) project conducted on CNBC news articles has provided valuable insights into the distribution and patterns of named entities in financial news. Through the analysis, we have identified significant trends and patterns in the occurrence of entities such as persons, organizations, and locations within the dataset. The visualizations generated have offered a clear representation of these findings, enhancing our understanding of market dynamics and the key players within the financial domain. The implications of this project extend to informed decision-making for stakeholders in finance, including investors, analysts, and policymakers, who can leverage these insights to make more informed decisions. Furthermore, the project has demonstrated the potential for developing specialized natural language processing (NLP) applications tailored to financial news analysis, with implications for improving information retrieval and decision support systems in finance. Overall, the NER project underscores the importance of NLP techniques in extracting actionable insights from unstructured textual data and highlights the value of NER in understanding and navigating complex financial markets.

## IX. References

*[1]. Perera, N., Dehmer, M., & Emmert-Streib, F. (2020). Named entity recognition and relation detection for biomedical information extraction. Frontiers in Cell and Developmental Biology, 8. https://doi.org/10.3389/fcell.2020.00673*

*[2]. A comprehensive guide to named entity recognition (NER). (2022, June 10). Turing.com; Turing Enterprises Inc. https://www.turing.com/kb/a-comprehensive-guide-to-named-entity-recognition*