Quantitative Performance Evaluation of the spaCy en_core_web_sm NER Model on the AG News Dataset

Juan Carlos Miguel Timoteo
Computer Studies & Engineering
José Rizal University
Mandaluyong City, Philippines
juancarlomiguel.timoteo@my.jru.edu

Abstract— **Named Entity Recognition (NER) is a foundational task in natural language processing for extracting structured information from text. This paper presents a detailed, step-by-step quantitative evaluation of spaCy's pre-trained en_core_web_sm model on the general news media domain. A rigorous methodology was employed, beginning with the curation of distinct text samples from the AG News dataset. A manual ground truth was then generated by annotating entities according to the OntoNotes 5.0 schema. These samples were subsequently processed by the spaCy model, and its automated output was compared against the ground truth. Precision, recall, and F1-score—standard metrics—were used to gauge performance. With an F1-Score of 0.700, the model showed a baseline performance that was moderately effective. The results analysis provides a clear baseline for the model's out-of-the-box capabilities by highlighting its proficiency in detecting common entity types while also exposing specific limitations in handling less common named entities, ambiguous labels, and exact entity bounds.**

Keywords—spaCy, en_core_web_sm, NER, Quantitative Analysis, News Text,

I. Introduction

Finding and categorizing named things in unstructured text into pre-established groups, such as people, organizations, and places, is the goal of Named Entity Recognition (NER), a subtask of information extraction [1]. For many sophisticated natural language processing (NLP) systems, NER is an essential front-end component that converts unstructured text into structured data.

The spaCy library is a widely adopted, open-source framework for industrial-strength NLP [2]. It provides a comprehensive suite of tools, including pre-trained statistical models. This study focuses on its popular en_core_web_sm model, a convolutional neural network trained on a large corpus of web text. While powerful, the performance of such general-purpose models can vary significantly depending on the specific domain of the text being analyzed. The objective of this paper is to quantitatively benchmark the baseline performance of this model on the highly relevant domain of news media, without any domain-specific fine-tuning.

To achieve this, the AG News Classification Dataset was selected as a representative testbed [3]. This paper meticulously documents the five-step methodology used, presents the quantitative results, discusses the specific patterns of success and error observed in the model's output, and concludes with an overall assessment.

II. Methodology

Before The evaluation was conducted following a structured five-step process designed to ensure replicability and robust analysis. The evaluation was conducted following a structured five-step process, detailed below, to ensure clarity and replicability.

A. Domain Selection

The domain of general news media was selected for this evaluation. This domain is highly suitable for NER tasks due to its high density and variety of standard named entities. The AG News dataset provides a large, clean, and well-structured collection of such text, making it an ideal benchmark.

B. Data Curation: The Selection of Three Samples

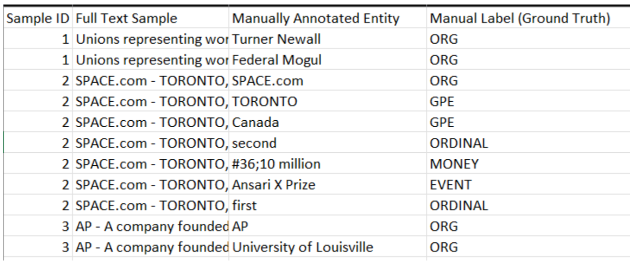
The train.csv file from the AG News dataset was used as the source for data curation. To create a small but diverse corpus for deep analysis, three distinct text samples were manually selected from the 'Description' column of the dataset. A diversity strategy was employed by selecting one sample from each of the 'Business' (Class 3) and 'Sci/Tech' (Class 4) categories. The three samples chosen were:

1. (Business) "Unions representing workers at Turner Newall say they are 'disappointed' after talks with stricken parent firm Federal Mogul."
2. (Sci/Tech) "SPACE.com - TORONTO, Canada -- A second team of rocketeers competing for the #36;10 million Ansari X Prize, a contest for privately funded suborbital space flight, has officially announced the first launch date for its manned rocket."
3. (Sci/Tech) "AP - A company founded by a chemistry researcher at the University of Louisville won a grant to develop a method of producing better peptides, which are short chains of amino acids, the building blocks of proteins."

C. Ground Truth Generation: Manual Annotation

To create a "gold standard" for evaluation, a manual annotation process was undertaken. The entities in the three samples were identified and labeled according to the entity schema used to train the spaCy model, which is based on the OntoNotes 5.0 release [4]. This schema includes 18 entity types, such as ORG (Organization), GPE (Geopolitical Entity), EVENT, MONEY, and ORDINAL.

An Excel spreadsheet was designed as an annotation tool, with columns for Sample ID, Full Text Sample, Manually Annotated Entity, and Manual Label. Each entity was meticulously recorded, resulting in a ground truth corpus of 11 named entities as shown in fig 1.



Sample ID	Full Text Sample	Manually Annotated Entity	Manual Label (Ground Truth)
1	Unions representing wor	Turner Newall	ORG
1	Unions representing wor	Federal Mogul	ORG
2	SPACE.com - TORONTO,	SPACE.com	ORG
2	SPACE.com - TORONTO,	TORONTO	GPE
2	SPACE.com - TORONTO,	Canada	GPE
2	SPACE.com - TORONTO,	second	ORDINAL
2	SPACE.com - TORONTO,	#36;10 million	MONEY
2	SPACE.com - TORONTO,	Ansari X Prize	EVENT
2	SPACE.com - TORONTO,	first	ORDINAL
3	AP - A company founded	AP	ORG
3	AP - A company founded	University of Louisville	ORG

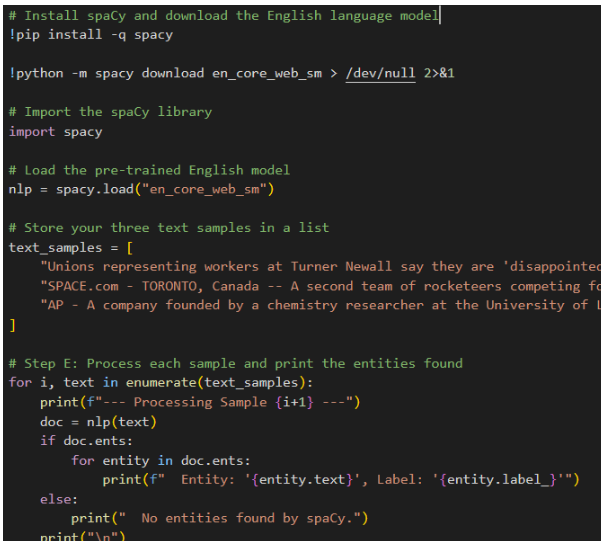
Fig 1. Output of spaCy Implementation

D. Automated Entity Recognition: Processing with spaCy

This step involved using the model being evaluated. The en_core_web_sm model from spaCy was used. The name of this model signifies:

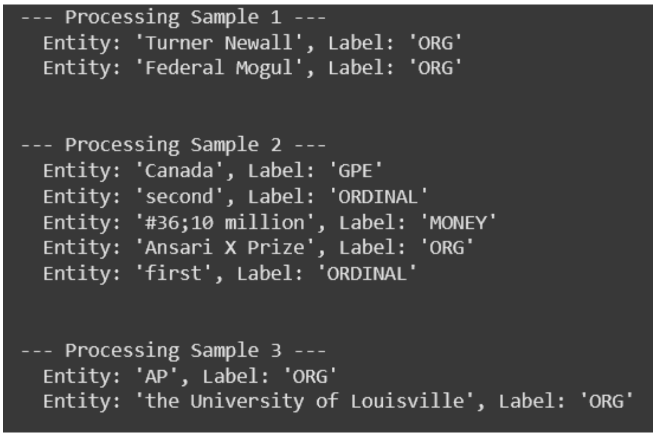
- **en**: It is an English language model.
- **core**: It provides core NLP capabilities (tagging, parsing, NER).
- **web**: It was trained on a large corpus of web text.
- **sm**: It is the small-sized model variant, optimized for efficiency.

The three text samples were processed programmatically using the spaCy v3 library in a Google Colab Python environment as shown in figure 2. The entities identified by the en_core_web_sm model, including their exact text and predicted labels, were extracted and recorded in the Excel template for a direct, side-by-side comparison with the ground truth. Output is shown in figure 3. And comparison is shown on figure 4.



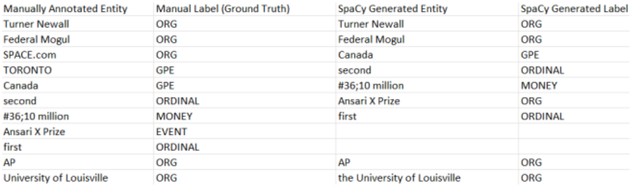
```
# Install spaCy and download the English language model
!pip install -q spacy

!python -m spacy download en_core_web_sm > /dev/null 2>&1

# Import the spaCy library
import spacy

# Load the pre-trained English model
nlp = spacy.load("en_core_web_sm")

# Store your three text samples in a list
text_samples = [
    "Unions representing workers at Turner Newall say they are 'disappointed
    "SPACE.com - TORONTO, Canada -- A second team of rocketeers competing fo
    "AP - A company founded by a chemistry researcher at the University of L
]

# Step E: Process each sample and print the entities found
for i, text in enumerate(text_samples):
    print(f"--- Processing Sample {i+1} ---")
    doc = nlp(text)
    if doc.ents:
        for entity in doc.ents:
            print(f"  Entity: '{entity.text}', Label: '{entity.label_}'")
    else:
        print("  No entities found by spaCy.")
    print("\n")
```

Fig 2. spaCy in Google Collab



```
--- Processing Sample 1 ---
  Entity: 'Turner Newall', Label: 'ORG'
  Entity: 'Federal Mogul', Label: 'ORG'


--- Processing Sample 2 ---
  Entity: 'Canada', Label: 'GPE'
  Entity: 'second', Label: 'ORDINAL'
  Entity: '#36;10 million', Label: 'MONEY'
  Entity: 'Ansari X Prize', Label: 'ORG'
  Entity: 'first', Label: 'ORDINAL'


--- Processing Sample 3 ---
  Entity: 'AP', Label: 'ORG'
  Entity: 'the University of Louisville', Label: 'ORG'
```

Fig 3. Output of spaCy Implementation



Manually Annotated Entity	Manual Label (Ground Truth)	SpaCy Generated Entity	SpaCy Generated Label
Turner Newall	ORG	Turner Newall	ORG
Federal Mogul	ORG	Federal Mogul	ORG
SPACE.com	ORG	Canada	GPE
TORONTO	GPE	second	ORDINAL
Canada	GPE	#36;10 million	MONEY
second	ORDINAL	Ansari X Prize	ORG
#36;10 million	MONEY	first	ORDINAL
Ansari X Prize	EVENT		
first	ORDINAL		
AP	ORG	AP	ORG
University of Louisville	ORG	the University of Louisville	ORG

Fig 4. Manual Label and Spacy Output Comparison

E. Performance Analysis and Accuracy Assessment

The model's performance was quantitatively assessed by comparing its predictions to the ground truth. A strict evaluation was used, where a prediction was counted as a True Positive (TP) only if its text boundary and label were an exact match. Predictions not present in the ground truth were counted as False Positives (FP), and ground truth entities missed by the model were counted as False Negatives (FN). The standard metrics were then calculated:

$$Precision = TP / (TP + FP)$$
$$Recall = TP / (TP + FN)$$
$$F1\text{-}Score = 2 * (Precision * Recall) / (Precision + Recall)$$

III. Results

The evaluation was performed on the 11 manually annotated ground truth entities. The en_core_web_sm model identified a total of 9 entities in the same text. The strict comparison yielded 7 True Positives, 2 False Positives, and 4 False Negatives. The final calculated performance metrics are presented in Table I.

Metric	Score
Precision	0.778
Recall	0.636
F1-Score	0.700

Table 1. Metric evaluation

A visual breakdown of the counts of True Positives, False Positives, and False Negatives is provided in Fig. 4. label, present them within parentheses.

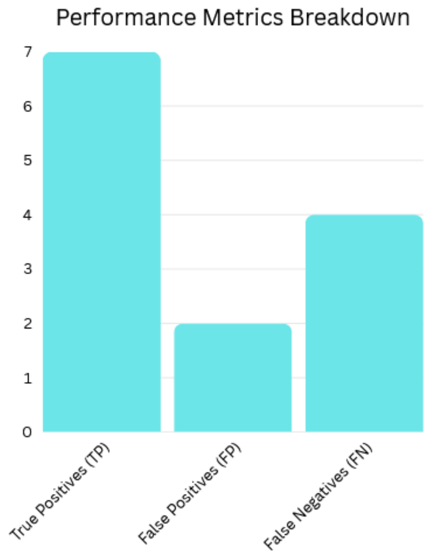


Fig 4. Performance Metrics Breakdown

This The results, culminating in an F1-Score of 0.700, indicate that the en_core_web_sm model is moderately effective for NER on news text out-of-the-box. An F1-Score of 70% is a respectable baseline, but a detailed analysis of the errors provides deeper insights.

The model demonstrated high proficiency in identifying common and unambiguously formatted entities. For example, it flawlessly identified Turner Newall and Federal Mogul as ORG entities. It also correctly tagged Canada (GPE), ordinal numbers, and monetary values. The successful identification of AP (Associated Press) as an ORG further showcases its robust training on common abbreviations found in news text.

However, the analysis revealed several key weaknesses:

Missed Entities (False Negatives): The model completely failed to identify SPACE.com and TORONTO. The former may be due to its unusual capitalization and domain-like structure, which falls outside the patterns learned from its web training data. The omission of TORONTO is a more surprising failure for a common GPE.

1. Misclassification: The most notable error was the mislabeling of Ansari X Prize as an ORG instead of an EVENT. This suggests that while the model recognized it as a named entity, its training was insufficient to distinguish this specific type of event name from a corporate name, despite EVENT being a valid label in the OntoNotes schema.
2. Boundary Errors: A subtle but important error occurred with the University of Louisville. The model incorrectly included the leading article "the" in the entity span. This highlights the challenge of determining precise entity boundaries, a common issue for token-based NER systems.
3. A breakdown of the error types is visualized in Fig. 3. These findings suggest the model's general-purpose training serves it well for high-frequency entities but falters when faced with less common proper nouns or nuanced classifications.

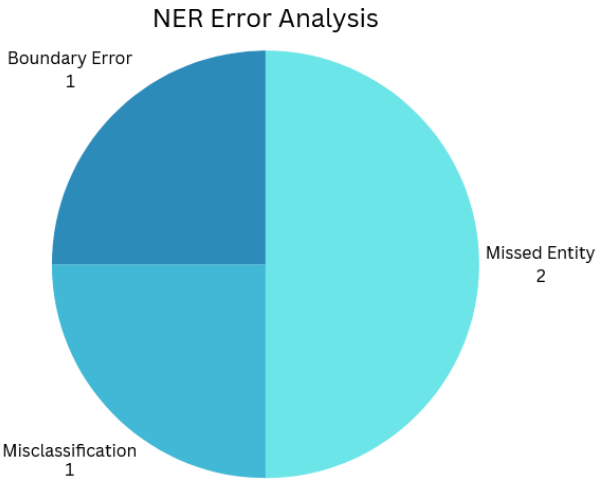


Fig 5. NER error Analysis

IV. CONCLUSION

This study conducted a detailed, step-by-step evaluation of the spaCy en_core_web_sm model's NER performance on the AG News dataset. The model achieved a final F1-Score of 0.700, establishing a clear performance benchmark.

The key finding is that while this pre-trained model is a powerful baseline tool, its reliability is highest for conventional, high-frequency entities. Its performance diminishes when faced with domain-specific names, unusual formatting, and the need for precise entity boundary detection. For applications requiring higher accuracy, these results strongly suggest the need for further fine-tuning. Future work should focus on training the spaCy model on a larger, custom-annotated corpus of news articles to improve its accuracy on these identified areas of weakness, thereby creating a more robust, domain-specific NER solution.

V. REFERENCES

[1] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Linguisticae Investigationes*, vol. 30, no. 1, pp. 3-26, 2007.

[2] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength Natural Language Processing in Python," 2020. [Online]. Available: https://spacy.io

[3] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed., Draft, Stanford University, 2023. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[4] A. Nandra, "AG News Classification Dataset," Kaggle, 2018. [Online]. Available: https://www.kaggle.com/datasets/amananandrai/ag-news-classification-dataset