Sydney Ani

Ekhtari

EDS 6397 – NLP

February 2, 2025

Named Entity Recognition Performance Assessment Using SpaCy

In this study, we will evaluate the performance and application of SpaCy’s Named Entity Recognition pipeline by comparing the automated entity tagging feature it holds against my manually annotated entity tagging. Our objective was to assess SpaCy’s accuracy in detecting and classifying named entities across 7 different entity groups: LOC, PERSON, NORP, ORG, GPE, DATE, and MONEY. The metrics used to test whether SpaCy was inclined in classifying the NER’s into these categories were Precision, Recall and F1 Scores. By conducting these tests, we were able to find that SpaCy’s NER pipeline compared to the Manually Annotated tweets revealed the best assessment was F1 Score.

A screenshot of a graph

Description automatically generated

Precision: Measures

correctness but isn’t

enough as this ignores

missing entities

Recall: Measures

completeness but isn’t

enough because this

ignores false positives

F1 Score: Balances both

completeness and

correctness, makes

best fit for eval.

As you can see from the photos illustrated above, F1 Score was the best performing assessment in terms of matching exactly what the Manual Annotations had. SpaCy showed a poor performance of 1/3 as for Recall and Precision, it overfit the data that it was given. Recall and Precision measure the completeness and correctness, SpaCy’s NER pipeline benefited off measuring both to ensure an overall performance thus we used F1 Score. Using F1 Score considered both false positives and false negatives, giving it a more harmonic approach and measure towards SpaCy’s NER accuracy.

Moving onto the Confusion Matrix shown below

A screenshot of a graph

Description automatically generated

The Confusion Matrix visualizes differences between my Manual Annotated tweets VS. the SpaCy’s NER Annotated tweets performance. We used the same exact metrics, Recall, Precision and F1 Score and you can see again the Positive (Red) values indicate that SpaCy outperformed manual annotations in that certain metric. The Negative (Blue) values indicate that SpaCy underperformed compared to manual annotations in that certain metric. And lastly, the Neutral (No color) values suggested a similar performance between SpaCy and the Manual annotations. ‘DATE’ shows the largest variation with both Precision and Recall while ‘NORP’ and ‘ORG’ show minor Recall improvement but slight Precision drops. I hope this helped visualize the over or under performing entities and provided insight into their strengths and weaknesses compared to the manual annotations.