

# ASSIGNMENT 1 NAMED ENTITY RECOGNITION

## REPORT

Name: Gopi Trinadh Maddikunta

PSID: 2409404

This report provides a comparison of accuracy between manually annotated and spacy annotated for the provided data of 300 tweets. I have manually annotated given data (300 tweets) by using of online NER annotation tool with 7 entities. They are PERSON, NORP, ORG, GPE, LOC, MONEY, DATE. I have take help of search engine to annotate manually likewise to differentiate between to GPE tag and LOC tag.

For phase 1, I have taken out all tweets and copied them into text file “Sentences.txt”, which are assigned to me. This text file loaded into **NER Annotator Tool**. Started annotated based on provided entities and exported a JSON file.

I have upload JSON file and make counter() function, which helps to count number of each tags available in the “annotations.json” file.

For phase 2, I have loaded a pretrained SpaCy NER model, en\_core\_web\_sm, which is used to extract Named Entities, identified based on entities.

Initially, The text file which consists of 300 tweets was read by using readlines() method, and stored in the list of strings in “lines” variable.

```
lines = f.readlines()
```

Secondly, Text input “Sentence” from list of strings “lines” were stripped and processed through the nlp() function and stored in a spacy object “doc”.

```
doc = nlp(sentence.strip())
```

1. sentence.strip() removes whitespace from beginning and ending of the string from variable “sentence”.
2. nlp() processes the passed cleared “sentence” to the nlp pipeline which has SpaCy pretrained model “en\_core\_web\_sm”, which helps to tokenizes and tagging of the text for Named Entity Recognition.
3. doc – It is SpaCy object which is used to stored the processed result into this variable. This SpaCy object consists of extracted named entities.

Subsequently, I have use the same counter() method which is used in the first phase were used as dictionary, stored keys as tags and values as Count of each tag, that helps to count the number of each entities present in the every tweet that has been provided.

For Phase 3, I have compared SpaCy's Predicted entities with the manual annotations to calculate the following metrics for each entity:

- **True Positives (TP):** The number of entities that were correctly predicted by the SpaCy.
- **False Positives (FP):** The number of entities predicted by SpaCy that were incorrect.
- **False Negatives (FN):** The number of entities in manual annotations that were not predicted by Spacy.

For each label, I have calculated Precision, Recall, F1 Score by using True Positives, False Positives and False Negatives counts:

- **Precision:** Measure how many predicted entities are correct.
- **Recall:** Measures of how many actual positives were predicted from all positives.
- **F1 Score:** It is a combination of Precision and Recall. Which is measure as Harmonic mean.

Classification Report:							
	Tag	TP	FP	FN	Precision	Recall	F1 Score
0	PERSON	81	18	15	0.818	0.844	0.831
1	NORP	85	20	11	0.810	0.885	0.846
2	ORG	81	32	21	0.717	0.794	0.753
3	GPE	177	38	23	0.823	0.885	0.853
4	LOC	26	7	38	0.788	0.406	0.536
5	DATE	137	36	3	0.792	0.979	0.875
6	MONEY	16	1	9	0.941	0.640	0.762

Fig: The Classification Report on performance of NER with comparison of SpaCy Predicted annotations with Manual annotations.

### Performance analysis of SpaCy's NER model:

**1. Precision:** A Precision score is high, it indicates that SpaCy's model predicts few false positives. What if Precision is low, it means SpaCy is incorrectly labeling entities where there are none.

“**MONEY**” tag has high Precision. So, there are very less False Positives predicted by SpaCy.

**2. Recall:** A high recall score indicates SpaCy was able to correctly identify most of entities present in the manually annotated data. A low recall indicates that SpaCy missed many entities.

“**LOC**” tag has low Recall. So, the SpaCy did not predict many LOC entities.

**3. F1 Score:** It reflects how well SpaCy's model is performing in terms of both recognizing entities correctly and covering all of them.

My SpaCy model well performed on annotating “**DATE**” tag which has high F1 Score.