Assignment 2.1: NLP, NER, and POS Tagging

Patricia Enrique

Department of Engineering, University of San Diego

AAI-511: Neural Networks and Learning

Prof. Mokhtari

September 18, 2023

Process

The dataset used for the named entity recognition (NER) and parts of speech (POS) tagging is retrieved from Kaggle and consists of 1535 claims dealing with climate-change collected from the internet (Ceunen, 2020). For this analysis, the column labeled claim is converted to a string and used. The first five claims from this dataset are summarized in Table 1.

Table 1

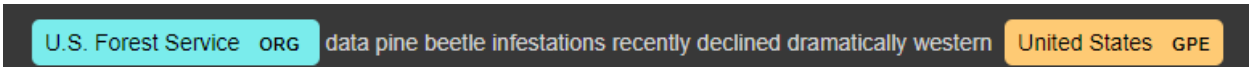
Climate change claims from the internet

ID	Claim
0	Global warming is driving polar bears toward extinction
5	The sun has gone into 'lockdown' which could cause freezing weather, earthquakes and famine, say scientists
6	The polar bear population has been growing.
10	Human additions of CO2 are in the margin of error of current measurements and the gradual increase in CO2 is mainly from oceans degassing as the planet slowly emerges from the last ice age.
1394	U.S. Forest Service data show pine beetle infestations have recently declined dramatically throughout the western United States.

The 640th claim, with ID 1394 as seen in Table 1, is used to perform and visualize NER and POS using the Spacy library before applying the techniques to the whole dataset. The singular claim is passed into the trained English pipeline and stop words and punctuation are removed from the text. When NER is applied to the text, U.S Forest Service is identified as an organization and United States is identified as a geopolitical entity as seen in figure 1.

Figure 1

NER application on claim ID 1394



Next, POS tagging is used to categorize each word in the text. This is completed with sequence classification which gathers context from surrounding words to better predict the part of speech of the current word (Bird et al., 2019). The selected claim contains five unique categorize which are summarized in Table 2.

Table 2

POS tagging application on claim ID 1394

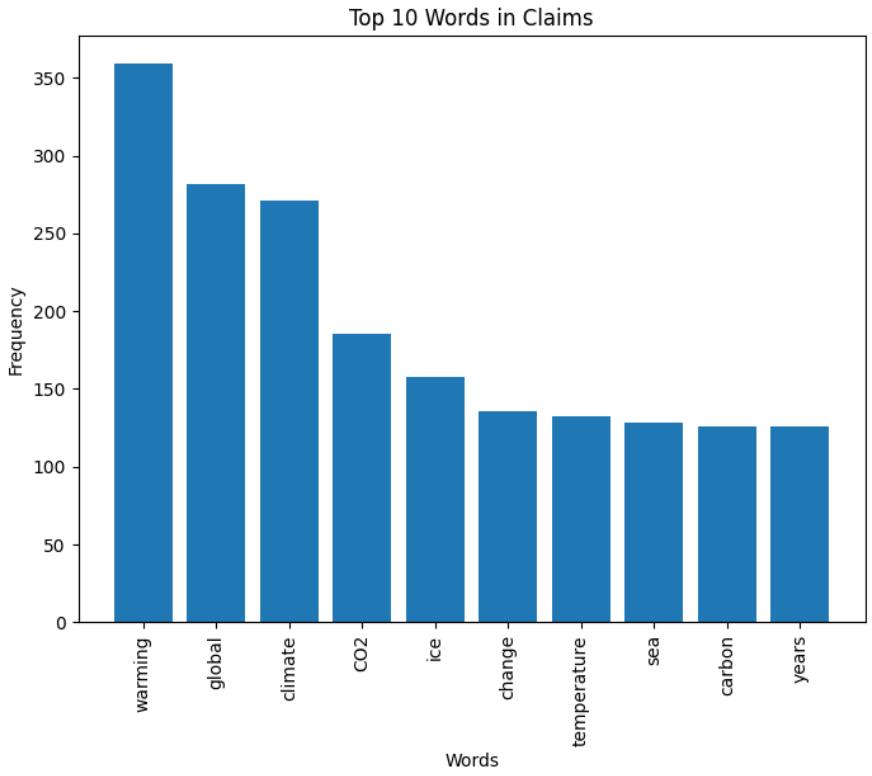
Claim 1394 Tokens	POS Tag
U.S	Proper noun
Forest	Proper noun
Service	Proper noun
data	Noun
pine	Noun
beetle	Noun
infestations	Noun
recently	Adverb
declined	Verb
dramatically	Adverb
western	Adjective
United	Proper noun
States	Proper noun

Results

When analyzing the complete dataset of claims, a similar approach is followed. The text is tokenized before the stop words and punctuation are removed, and the remaining text is analyzed. The frequency of the tokens in the text are illustrated in Figure 2. It can be seen that the words are mainly associated with the topic of climate change and provide insight on the recurring topics discussed in many of the claims.

Figure 2

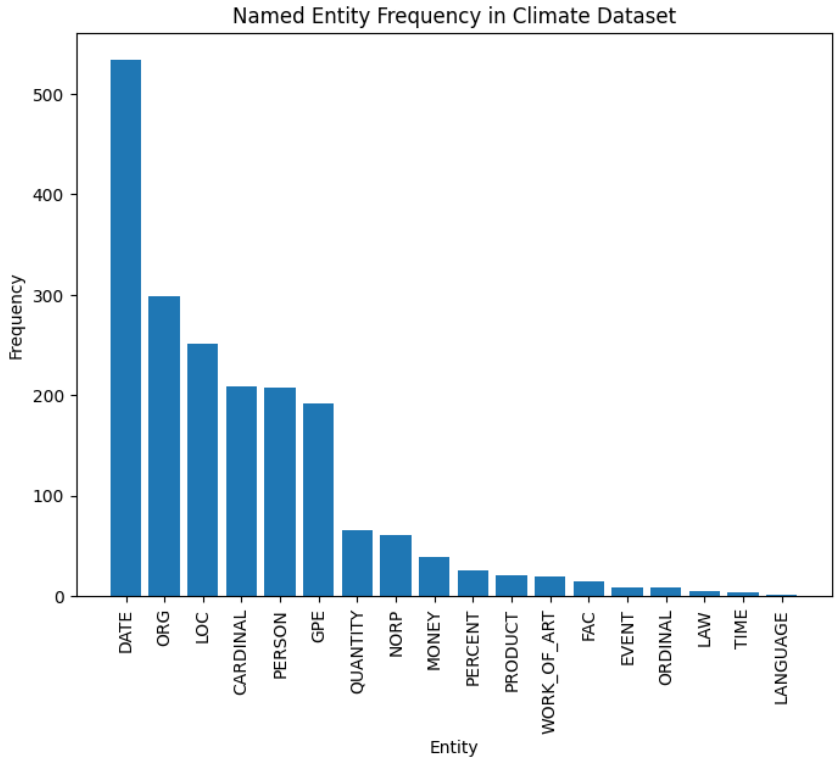
Top 10 words in the climate change dataset



The tokens extracted from the dataset are recombined into a string for NER and POS tagging. When NER is applied to the completed dataset of claims, the frequency of each entity is shown in Figure 3. It can be seen from this graph that organizations and people are commonly found in the claims; however, quantitative values such as percentages and laws are not as common. This can be an indicator that the claims are blaming organizations or making statements that are not factually backed or supported.

Figure 3

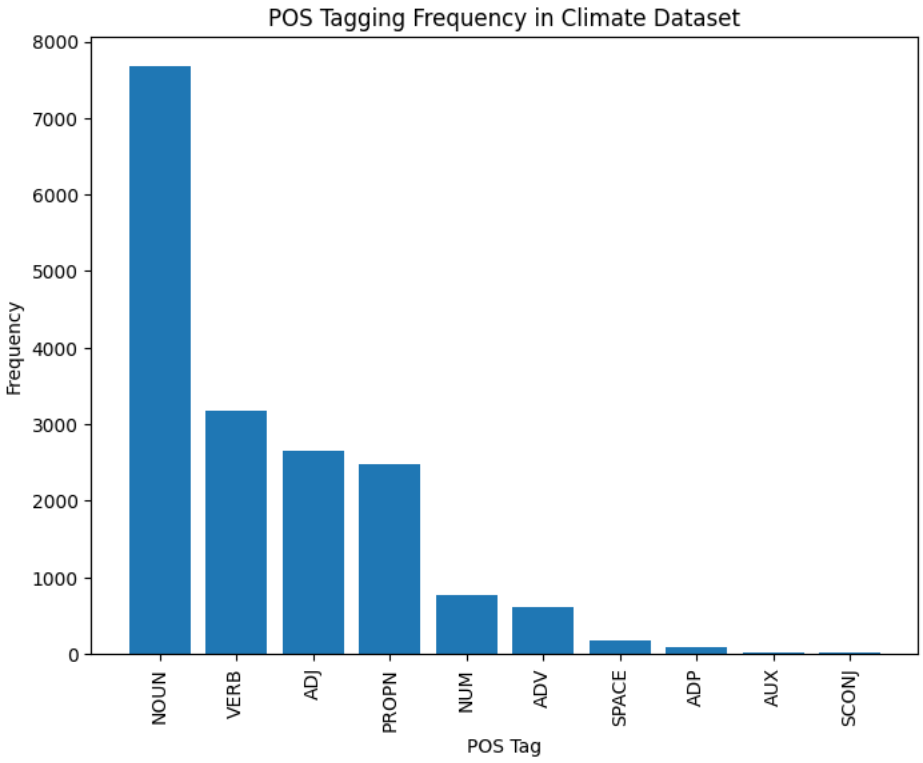
Frequency of each entity recognized in the dataset



When POS tagging is applied to the completed dataset of claims, the frequency of the top ten occurring categories are shown in Figure 4. It can be seen that nouns are used more than two times than any other POS. Majority of the text falls into one of six categories: noun, verb, adjective, proper noun, number, and adverb. This result supports the findings from the NER application, as the amount of numbers used in the claims is drastically smaller than the amount of nouns and proper nouns.

Figure 3

Frequency of the top 10 POS tagging in the dataset



Conclusion

A few challenges were faced when completing this analysis. It is observed that words such as show and throughout are considered to be stop words when using the Spacy library. For this analysis this distinction is not critical; however, depending on the application and the context, the stop words may need to be reviewed and certain ones kept in the text instead of being omitted. For NER and POS tagging techniques which rely on contextual information, a bag-of-words model is not sufficient. Instead, the tokens extracted from the text have to be rejoined to maintain the context. When completing POS tagging, it is noted that a category present is defined as space, which does not provide much information and is better off omitted for more valuable information. Challenges that were not encountered in this analysis but should be considered when completing NER and POS tagging tasks include word ambiguity and unknown words with inconsistent tags.

References

Bird, S., Klein, E., & Loper, E. (2019). Natural Language Processing with Python. *Natural Language Toolkit.* https://www.nltk.org/book/ch05.html

Ceunen, B. (2020). *Climate Fever Dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/bouweceunen/climate-fever-dataset

Li, S. (2018, August 17). Named Entity Recognition with NLTK and SpaCy. *Towards Data Science*. https://towardsdatascience.com/named-entity-recognition-with-nltk-and-spacy-8c4a7d88e7da