**ASSIGNMENT\_2\_AAI\_520\_NLP\_NER AND POS Tagging**

KWABENA MENSAH

*University of San Diego*

**Climate Fever Dataset**

**Content**

A dataset adopting the FEVER methodology that consists of 1535 real-world claims regarding climate-change collected on the internet. Each claim is accompanied by five manually annotated evidence sentences retrieved from the English Wikipedia that support, refute or do not give enough information to validate the claim totaling 7675 claim-evidence pairs. The dataset features challenging claims that relate multiple facets and disputed cases of claims where both supporting and refuting evidence are present.

**Text Preprocessing Approach**

Text preprocessing is a crucial step in Natural Language Processing (NLP) and Machine Learning tasks, ensuring that the input data is cleaned and standardized to improve the efficiency and accuracy of subsequent processes. This report summarizes the preprocessing approach employed:

**HTML Tag Removal:**

All HTML tags present in the text are removed to ensure the extraction of pure content without any formatting instructions.

**Contraction Expansion:**

Contractions are expanded to their full form. For instance, "I'm" becomes "I am".

**Number Removal:**

All numeric values present in the text are removed to focus solely on textual content.

**URL Removal:**

Any URLs present in the text are removed, ensuring that external references don't influence the analysis.

**Mentions Removal:**

Mentions, often found in tweets, and represented by '@', are removed to keep the focus on the content and not on specific users.

**Tokenization:**

The text is broken down into individual words or tokens. This aids in analyzing and processing the text at a granular level.

**Stopword Removal:**

Common words that do not contribute significant meaning to the text, known as stopwords, are removed. Examples include "and", "the", and "is".

**Punctuation Removal:**

All punctuation marks are removed from the text to ensure a smooth flow and reduce noise.

**Non-ASCII Character Removal:**

Characters that are not part of the standard ASCII set are removed to maintain uniformity and compatibility.

**Hashtag Removal:**

Hashtags, which are often used in social media posts, are removed to avoid potential biases in analysis.

**Lemmatization:**

Words are reduced to their base or dictionary form. For example, "running" becomes "run". This helps in standardizing words and reducing the overall vocabulary.

In conclusion, while all the preprocessing steps outlined above may not have been necessary, it ensures that the data is comprehensively cleaned and the described preprocessing approach provides a comprehensive cleaning and standardizing methodology for text data, making it more amenable for further analysis and modeling. This a pipeline created so that I do not have to spend time reviewing text data to determine types of cleaning steps required.

**Tools Used in Text Preprocessing and Analysis:**

Spacy: Employed for lemmatization and stopwords removal. It's a comprehensive library offering various NLP functionalities.

NLTK: Used for tokenization of words. It is one of the most popular libraries for natural language processing tasks.

TextBlob: Is a Python library for processing textual data, offering a simple API for diving into common NLP tasks. Among its many functionalities, TextBlob can be used for obtaining word counts and frequencies in a text.

NER and POS Tagging

NER refers to the process of classifying named entities in text into predefined categories such as persons, organizations, locations, etc.

***Part-of-Speech (POS) Tagging was used for the following reasons:***

1. Information Extraction: Helps in extracting structured information from unstructured data sources, like news articles or web pages, facilitating data organization and retrieval.
2. Enhancing Search Engines: Allows search engines to understand the significance of named entities in a document, enabling more context-aware search results.
3. Content Recommendation: By understanding named entities, systems can recommend content more relevant to a user's interests.
4. Relation Extraction: Helps in understanding the relations between different named entities in the text, aiding in knowledge graph construction and semantic web applications.

POS Tagging involves labeling each word in a text according to its corresponding part of speech (e.g., noun, verb, adjective).

***Part-of-Speech (POS) Tagging was used for the following reasons:***

1. Disambiguation: Helps in resolving the meaning of a word based on its use in a sentence. For example, distinguishing between "book" (noun) as in "I read a book" and "book" (verb) as in "I will book a ticket."
2. Syntax Parsing: It serves as a foundational step for parsing the structure of sentences, enabling the extraction of relationships and dependencies between words.
3. Lemmatization: Assists lemmatizers in determining the base form of words. For instance, knowing that "running" is a verb helps in lemmatizing it to "run" instead of the noun form "running."
4. Enhancing Machine Learning Models: POS tags can be used as features in machine learning models, particularly in tasks like sentiment analysis where the adjective-noun structure might be significant.

In performing NER and POS, I used the two columns **claim** being the independent variable and claim\_label the dependent variable. From reading the requirements of the assignment adding the other columns from the evidence will have made it more complex than is maybe required for this assignment. Additionally, it appears the evidences column provides information about the evidence used to determine the veracity of the claim which is our independent column.

**Challenges:**

The main challenge was that the data set was small, just 1535 rows.

**Observations/Findings:**

Using the SUPPORTS and DISPUTES categories as illustration in this report, we noted that performing NER and POS after text preprocessing, facilitates a deeper understanding of the context and structure of sentences, enabling us to better understand the meaning and setup of sentences and most importantly words, leading to improved uses in Natural Language Processing. NER and POS Tagging are key tools that help us get detailed and specific information from text.

**Visualizing the DataFrame Before Text Prepocessing - SUPPORTS**

A graph of a number of words

Description automatically generated with medium confidence

**Visualizing the DataFrame After Text Prepocessing - SUPPORTS**

A graph with blue and white bars

Description automatically generated

**Visualizing the DataFrame Before Text Prepocessing – DISPUTED**

**A graph of a number of words

Description automatically generated**

**Visualizing the DataFrame After Text Prepocessing – DISPUTED**

**A graph of a number of words

Description automatically generated**