**Code Book and Methodology**

**Introduction**

The present methodology provides an easily-accessible dataset to permit easier analysis of tweets authored by members of the U.S. Congress, spanning the period from June 2017 to July 2023. These tweets were collected from the "Tweets of Congress" repository available on GitHub.[[1]](#footnote-1) The primary objective of this project is to establish a robust Natural Language Processing (NLP) pipeline that effectively preprocesses and prepares this corpus of tweets for subsequent NLP analyses.

By constructing a streamlined NLP pipeline, the aim is to facilitate ease of access and usability for researchers seeking to investigate the linguistic nuances, sentiment dynamics, and thematic patterns in congressional communication. The processed dataset is intended to serve as a foundational resource for a diverse range of research inquiries, from examining the impact of political events on discourse to investigating sentiment trends across legislative sessions.

In pursuit of this goal, the NLP pipeline encompasses several critical steps. The initial stage involves text preprocessing, which encompasses techniques such as the removal of stop words, stemming, and lemmatization. These procedures help standardize the text data, enhancing the accuracy of subsequent analyses. The subsequent phase leverages advanced NLP tools, including the SpaCy language model, for Named Entity Recognition (NER). This extraction process identifies entities such as names, organizations, and locations, enriching the dataset with structured information.

Furthermore, sentiment analysis constitutes a pivotal component of the NLP pipeline. The Sentiment Intensity Analyzer assigns sentiment scores to extracted entities, enabling a quantifiable assessment of the emotional tone expressed in congressional communication. By combining these components, the processed dataset offers a multifaceted perspective on the language, sentiment, and entities embedded within congressional tweets.

In summary, this research initiative strives to establish a refined corpus of congressional tweets, characterized by its accessibility, accuracy, and richness. The resulting NLP pipeline empowers researchers to embark on a variety of investigations into the linguistic trends and sentiment dynamics of U.S. congressional communication. By providing this resource, we aspire to contribute to the scholarly exploration of the language utilized by the U.S. Congress, shedding light on its evolving discourse in the digital age.

**Libraries and Dependencies**

To see the dependencies, please review requirements.txt

**Data Collection and Preparation**

* The corpus of data comes from a GitHub repository[[2]](#footnote-2), which pulled tweets from twitter. The data comes as JSON files within zip archives.
* Please note that while the dataset captures data for nearly all days, there are instances where certain days are missing.

**Text Preprocessing**

Stop words Removal

* We used the stop words from the NLTK corpus, but added additional characters to be removed. Additional characters removed (csv):



* Excluding common words known as "stop words" from text analysis enhances the accuracy by focusing on meaningful content, while also improving computational efficiency. This strategy ensures that the analysis highlights essential words and themes within the text, leading to more precise insights.

Stemming and Lemmatization

* Stemming and lemmatization are pivotal techniques employed to enhance the accuracy and consistency of text analysis. Both processes aim to reduce inflected or derived words to their base or root form, thereby reducing variation and allowing for more effective comparison of words.
* We chose the Porter Stemmer, a widely-used algorithm known for its efficiency in truncating words to their root. It employs a series of rules to remove common prefixes and suffixes, producing a simplified version of the word.
* WordNet Lemmatizer is the chosen lemmatization tool. It employs a more sophisticated approach by utilizing lexical resources to transform words into their canonical form, or lemma.
* Example, Word: Running
  + Stemming: Run
  + Lemmatization: Run

**Natural Language Processing (NLP)**

NER and SpaCy

* Named Entity Recognition (NER) is a fundamental technique in natural language processing that involves identifying and categorizing entities within text, such as names of people and organizations.
* In this project, the **en\_core\_web\_sm** model provided by SpaCy assumes a central role in executing NER on processed text data. With the assistance of SpaCy's advanced language model, the text undergoes a comprehensive analysis, effectively identifying and classifying entities with exceptional accuracy.
* This results in the creation of a tagged representation of the text, accentuating critical elements that hold substantive meaning within the context

**Sentiment Analysis**

Sentiment Intensity Analyzer

* Sentiment Analysis is a powerful method employed to assess and quantify the sentiment or emotional tone present within text. This technique serves as a valuable tool to understand the prevailing sentiment, be it positive, negative, or neutral, encapsulated within textual content.
* In this project, the Sentiment Intensity Analyzer a component provided by the Natural Language Toolkit (NLTK), evaluates the sentiment of a given text and produces a sentiment score that reflects the overall emotional orientation of the content.
* The sentiment score assigned by the Sentiment Intensity Analyzer includes negative, neutral, and positive scores, ranging from -1 to 1.

Sentiment Analysis Process

• Sentiment analysis was employed on entities identified through NER, assessing their sentiment.

• The compound sentiment score was utilized to evaluate the overall sentiment of entities, ranging from negative to positive.

**Summary and Conclusion**

In this methodology, we embarked on a comprehensive journey of text analysis to unveil insights from the corpus of US congressional tweets. Key steps included data preprocessing, which involved cleaning text and eliminating stop words. Additionally, stemming and lemmatization normalized words, while Named Entity Recognition (NER) identified entities for further analysis. The powerful SpaCy language model was harnessed for NER and subsequently, Sentiment Analysis provided a nuanced understanding of sentiment within the text. The sentiment scores enabled the assessment of entities' emotional tone.

Ultimately, this methodology's integration of advanced NLP techniques not only facilitated data preparation but also empowered us to derive valuable insights from unstructured text. By unveiling sentiment, identifying entities, and understanding their emotional context, we transcend mere data analysis and tap into the essence of the language used by the US Congress. This process, informed by robust tools and methods, underscores the methodology's significance in generating meaningful insights from the text data, contributing to enriched research and understanding.

**Code Availability**

Access to code available through the Github repository:

<https://github.com/Chasen-Jeffries/US_Congress_Tweet_NLP_Dataset>

**Limitations and Future Work**

1. **Selective Stop Words**: Deliberate reintroduction of specific stop words, especially those connected with numbers or hashtags, could be considered to evaluate their influence on analysis results.
2. **Emojis as Words**: The possibility of deconstructing emojis into textual representations might be explored to enhance sentiment analysis accuracy and capture nuanced emotions.
3. **Gold Standard for NER**: Developing a gold standard dedicated to Named Entity Recognition (NER) on politicians’ tweets could offer improved model training and evaluation, potentially elevating the precision of entity identification.

1. <https://github.com/alexlitel/congresstweets> [↑](#footnote-ref-1)
2. <https://github.com/alexlitel/congresstweets> [↑](#footnote-ref-2)