**US Congressional Tweets NLP**

**Dataset Codebook**

**By: Chasen Jeffries**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

This dataset delves into a curated selection of tweets from the US Congress, spanning the period of June 2017 to July 2023. The dataset comprises a total of [total observations] entries, with a focus on individual congressional members as the primary units of analysis. Notably, the dataset is meticulously organized by day, month, and year, offering a granular perspective on the temporal dynamics of congressional communication. The core dataset encompasses information on tweet authors, posting dates, times of day, tweet content, preprocessed text, tweet statistics, named entities recognized by NER, and sentiment scores associated with those entities. This comprehensive dataset serves as a robust foundation for uncovering linguistic patterns, emotional nuances, and communication trends within the realm of the US Congress.

**Citations:**

Litel, A. (2023, July 1). congresstweets. GitHub. <https://github.com/alexlitel/congresstweets>

**Variables:**

**username**: The username of the tweet author.

RepErikPaulsen

**date**: The date when the tweet was posted.

6/21/2017

**time**: The time of day when the tweet was posted.

10:05:17 AM

**full\_tweet\_text**: The complete text content of the tweet.

RT @CRN\_Supplements Thank you @RepErikPaulsen for speaking to the dietary supplement industry today! And thank you for your support of HR 1175! #DoH17 #HSAs https://t.co/oADXe2kFR6

**cleaned\_text**: The preprocessed and cleaned version of the tweet text.

rt crn\_supplement thank reperikpaulsen speak dietari supplement industri today thank support hr 1175 doh17 hsa http //t.co/oadxe2kfr6

**num\_tweets**: The number of tweets by the same user on the same date.

6

**num\_chars**: The character count of the tweet text.

180

**num\_words**: The word count of the tweet text.

25

**num\_sents**: The sentence count in the tweet text.

2

**num\_vocab**: The count of unique words in the tweet text.

22

**ner\_entities**: Entities identified by Named Entity Recognition (NER).

['today', '1175', 'doh17']

**entity\_sentiment\_scores**: Sentiment scores for entities identified by NER.

[0.0, 0.0, 0.0]