1 Text Pre-processing

The `clean_tweets` function is designed to preprocess text data, specifically tweets, by performing several cleaning steps. First, it removes URLs, retweet markers (e.g., "RT @username:"), and mentions (e.g., "@user"). It then eliminates hashtags, replaces the "&" character reference with the word "and", and removes punctuation marks. Newline characters are replaced with spaces, and the entire text is converted to lowercase. Extra whitespace at the beginning and end of the text is trimmed. The function also removes stop words, defined by a custom regular expression (`stop_words_regex`), and eliminates any numbers from the text. This process results in cleaner, standardized text suitable for analysis or modeling.

2 Ideological Scoring

The method is taken from Le Pennec (2024) and Di Tella et al. (2023) and builds on Laver et al. (2003). The method is a supervised approach to scale each tweet on a Democrat-Republican scale. Since we know the party affiliation of each politician, we can infer the party affiliation of each tweet. This allows me to determine the political orientation of individual words, resulting in a more precise estimate of the tweet's political alignment based on word choice. The ideological score of tweet j is defined as:

$$S_j = \sum_w p_{wj}.s_w \tag{1}$$

Where p_{wj} is the frequency of the word w in tweet j defined as $p_{wj} = \frac{c_{wj}}{m_j}$ with c_{wj}, the number of occurrences of word w in tweet j and m_j, the total number of words in tweet j. Furthermore, s_w is the word score of word w which is defined in the following way:

$$s_w = \frac{p_w^R - p_w^D}{p_w^R + p_w^D} \tag{2}$$

Where $p_w^I = \frac{1}{|I|} \sum_{j \in I} p_{wj}$ is the average frequency of word w among documents from ideological side being $I \in \{R, D\}$ with R standing for Republicans and D for Democrats. The score of each word ranges between -1 and 1, -1 being for words that were only employed by Democrats, and 1 being for words that were only employed by Republicans. I remove words which are too (in)frequently used and I normalize the score as proposed by Martin & Vanberg (2008) Naturally, a tweet with a negative score contains words that were mainly used by Democrats while a tweet with a positive score contains words that were

mainly used by Republicans. Finally, a tweet with a score close to 0 contains words that were used equally by Democratic politicians and Republican politicians.

3 Sentiment Analysis

For the sentiment analysis, I use the `syuzhet` package. This **R** package is specifically designed for text sentiment analysis. The Syuzhet method assigns a sentiment score to text, where positive scores indicate a favorable or happy sentiment, and negative scores indicate an unfavorable or sad sentiment. The score is calculated based on a sentiment dictionary, with higher values reflecting stronger positivity and lower values indicating stronger negativity.You find find more details on syuzhet GitHub Repository

4 Complexity Measure

The Flesch-Kincaid Grade Level index (Flesch, 1948) measures the readability of a text based on sentence length and word complexity. It estimates the U.S. school grade required to understand the text, with lower scores indicating easier readability and higher scores signifying greater complexity. The formula for the index is:

$$\text{Grade Level} = (0.39 \times \frac{\text{Total Words}}{\text{Total Sentences}}) + (11.8 \times \frac{\text{Total Syllables}}{\text{Total Words}}) - 15.59$$

A higher value suggests a more complex text, while a lower value means the text is easier to comprehend. I compute this measures using the `quanteda.textstats` package.

References

Di Tella, R., Kotti, R., Le Pennec, C., & Pons, V. (2023). *Keep your Enemies Closer: Strategic Platform Adjustments during US and French Elections*. Technical report, National Bureau of Economic Research.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3), 221–233.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American political science review*, 97(2), 311–331.

Le Pennec, C. (2024). Strategic campaign communication: Evidence from 30,000 candidate manifestos. *The Economic Journal*, 134(658), 785–810.

Martin, L. W. & Vanberg, G. (2008). A robust transformation procedure for interpreting political text. *Political Analysis*, 16(1), 93–100.