NLP Analysis of Inaugural and State of the Union Discourse

Kevin Huang
Emory University
kevin.huang@emory.edu

Shuyang Bian
Emory University
simon.bian@emory.edu

Abstract

The United States presidential inaugural addresses and State of the Union speeches (SOTU) have their unique political significance—the former being a blueprint that presents the President's vision of the future of the States, and the latter a summary of the country's affairs up to that point. The consistency in occurrence and intent of these corpora may create a useful set of data points spanning the country's history. Using a Natural-language-Processing (NLP) approach, these centuries-spanning trends might be identified. Sentence complexity and document length suggest a gradual decrease in length and increase in readability over the years. N-grams analysis revealed a potential involvement of the religiosity of the United States and the role it plays in government. Using Gephi and network analysis, we revealed that presidents cite past presidents to form some cliques, which suggest connected political beliefs over time. Using BERT as a sentiment analysis model, we revealed an increased positivism in the speech over the years. Juxtaposing the GIS results, we further pinpointed important trends related to two types of documents: the NER data shows that 47% of locations mapped were international, whereas only 21% of locations in the INA corpus were international.

1 Introduction

Both US presidential inaugural addresses and State of the Union speeches have been heavily analyzed over the years, using both qualitative and computational methods. Generally, these analyses follow a common framework by which researchers attempt to draw some kind of insight into the political atmosphere, presidential language and goals of politicians, or underlying historical trends by using various tools in NLP or other methodologies.

2 Data and Methods

Both the inaugural and State of the Union corpora were collected from the UCSB American Presidency Project (The American Presidency Project, n.d.) The inaugural corpus consists of speeches from 1789 (Washington) to 2020 (Biden), totaling 63 speeches. The State of the Union corpus contains 242 total speeches from 1790 to 2023. These corpora were collected from the UCSB American Presidency Project using an ad-hoc bs4 script in Python which is available on the NLP Suite (https://github.com/NLP-Suite/NLP-Suite/wiki).
We then divided the two corpora along three ideas: war/peacetime, economic growth/decline, and party affiliation. The data for wartime and peacetime is taken from a US Congressional Research Service analysis (Daggett, Stephen. (2010). Costs of Major U.S. Wars. 9).
The basic foundation of this sub-corpus is to compare speeches given during different wars and between war and peace time. An important observation to note is that there are sometimes unclear delineations between whether the country is at war or at peace. This sometimes results in a difficulty to clearly place when a war truly started. For example, the Vietnam war is cited to start and end from 1964-1975. However, tensions in this region existed since the end of World War II, and during the 1950's, stopping the advance of communism in Asia became an object of incredible importance to US presidents (Kissinger, 2003, pp. 17-22). By the end of the Eisenhower administration (1961), 1500 American personnel were in South Vietnam along with over $ 1 billion in aid (Kissinger, 2003, p. 26) Clearly, the exact demarcation of a war date would be insufficient in determining when exactly US military activity arose. In these cases, it is important to apply a critical eye to a contextual understanding of these dates in order to fully capture the concept of "at war." In this sense, the separation

of war and peacetime by year dates is more of a heuristic for war and peacetime rather than a specific delineation. In addition, casual observers will notice that the sub-corpus of peacetime is much smaller than wartime. This is of course a reflection of the fact that the United States has frequently and for most of its existence been involved in a war in some capacity. Therefore, the analyses used on the War and Peace sub-corpus must be independent of corpus size and normalized for the length and quantity of text.

3 Results

From the data, there is an observable decrease in the average grade level from 1790 to 2023: visually, speeches started around the "15th grade" and began declining gradually, having since fallen to around the 5-10th grades. Furthermore, average Yngve value per sentence in a speech has decreased from a score of around 200 in 1790 to 50 in 2023. This suggests that the average complexity of the speeches on a sentence level is also decreasing over time. Outside of these eight presidents, the only other presidents who have any mentions are Richard Nixon, James Monroe, Calvin Cooolidge, George Bush, Herbert Hoover, Benjamin Harrison, and James Monroe. In total, 15/45 presidents have been mentioned by their peers throughout time.
GIS mapping of locations between the speeches allow for the charting of locations on Google Earth maps, which visually represent the prevalence of various locations within inaugural or State of the Union speeches. This method shows that the State of the Union corpus is much more international than the inaugural speeches—presidents are more likely to speak about non-American locations in State of the Union speeches than inaugural speeches. Analysis of sentiment scores across both inaugural and State of the Union speeches indicates that neither unemployment nor inflation rates seem to correlate strongly with the sentiment of the corresponding speech. However, these indicators alone may not be sufficient to understand the political atmosphere surrounding the economy at the time the speeches were given, as well as the fact that other events will influence the sentimental outlook of each speech. As it relates to war influencing sentiment, sentiment analysis applied to major wars and "peaceful" years reveals that the least positive presidential sentiments were associated with the War of 1812, World War 1, and World War 2. However, the overall average sentiment across all wars was relatively similar to that of "peaceful" years, indicating limited fluctuations in sentiment during these conflicts.

Limitations

Over the development of the NLP Suite, much effort has been recently put into streamlining the research process. Cache, for instance, has been widely used to store temporary Stanza results for allowing better computation of repeatedly requested results. Nevertheless, we recognize that NLP Suite's environment requirement has been an ongoing one. We are currently reformatting the program to be a less challenging, more beginner-friendly web version. We also recognize that we have yet to compare the variation in English by itself, as most of the language models we have used, have been trained on some more recent corpora. Therefore, in future endeavors, we would focus on contextualizing the diachronic effects of training corpora's effect on the presentation of the results.

Acknowledgements

We appreciate Kevin Huang, Dr. Roberto Franzosi, Simon Bian, and all other LING 489 W participants during the time.

References

https://doi.org/10.1073/pnas.151222111. Accessed: 2024-2-7.

Olusegun Oladele Jegede. 2020. Syntactic analysis of donald trump's inaugural speech. *ELS Journal on Interdisciplinary Studies in Humanities*, 3(3):317–327.

Ping Li, Benjamin Schloss, and D Jake Follmer. 2017. Speaking two "languages" in america: A semantic space analysis of how presidential candidates and their supporters represent abstract political concepts differently. *Behav. Res. Methods*, 49(5):1668–1685.

Ryan Light. 2014. From words to networks and back. *Soc. Curr.*, 1(2):111–129.

Saif Shahin. 2016. When scale meets depth: Integrating natural language processing and textual analysis for studying digital corpora. *Commun. Methods Meas.*, 10(1):28–50.

A Appendix

You may visit the NLP Suite at: https://github.com/NLP-Suite/NLP-Suite. You may contact the primary author, Kevin Huang, for the original data.