How do Presidents Speak?

Hunter Dawley
Vanderbilt University
hunter.b.dawley@vanderbilt.edu

ABSTRACT

This study examines the transformative role of presidential rhetoric in shaping American political and social landscapes, leveraging both historical insights and modern analytical techniques. It emphasizes the evolution of presidential communication styles, noting a significant shift from analytic to assertive rhetoric over the last century, influenced by technological advancements and changing media dynamics. Central to this analysis is the investigation of whether a logistic regression model can accurately classify presidential speeches into their respective political parties based on textual features such as TF-IDF vectorization and semantic similarity. This research not only explores the distinct linguistic patterns that differentiate political parties but also seeks to understand how these patterns reflect broader shifts in political strategy and public engagement. By bridging historical rhetorical analysis with contemporary computational methods, this paper aims to provide a comprehensive understanding of the enduring impact of presidential speeches on American society and the potential for predictive analytics in political science.

Keywords

Semantic similarity, TF-IDF, political parties, presidential speech.

1. INTRODUCTION

Throughout U.S. history, the significance of presidential speeches has been profound. Richard Neustadt famously stated in 1960 that 'presidential power is the power to persuade,' a notion that resonates even more in today's era of widespread social media usage (Beasley, 2021). Presidential speeches are not merely reflections of current policies and ideologies; they are powerful tools for shaping public opinion and inspiring collective action. For instance, Lyndon B. Johnson's 'Great Society' speech, which outlined a vision to eradicate poverty and racial injustice, exemplifies how presidential addresses can set ambitious societal goals and mobilize public imagination and initiative to achieve them (Kibler, 2021). Similarly, his 'We Shall Overcome' speech became a cornerstone of the civil rights movement, using the presidential platform to amplify the call for justice and equality (Kibler, 2021). Ronald Reagan's speeches, such as the D-Day Anniversary Address and the Berlin Wall Speech, showcase the power of storytelling and historical references in reinforcing political messages. These speeches not only commemorated significant historical events but also revitalized political alliances and democratic ideals, demonstrating the impact of presidential rhetoric on national discourse during pivotal moments in history (Kibler, 2021). Moreover, recent discussions have highlighted how leaders like Donald Trump have used language to influence the country's state, with some commentators labeling such communications as atypical. However, linguistic analyses suggest that over the last century, there has been a consistent decline in analytic thinking and a rise in assertiveness in presidential communication styles, which significantly affect public engagement and perception (Pennebaker, 2019). The evolution of presidential speech trends—from Johnson's civil rights initiatives to Reagan's Cold War rhetoric—reflects shifts in political communication strategies and the dynamic interaction between leadership, public expectations, and media developments. Studying these trends offers invaluable insights into the changing priorities and challenges of different eras, underscoring the enduring influence of the presidential office in shaping the American narrative. The study of presidential speeches is crucial as it affects every individual in society and has the potential to drive significant change.

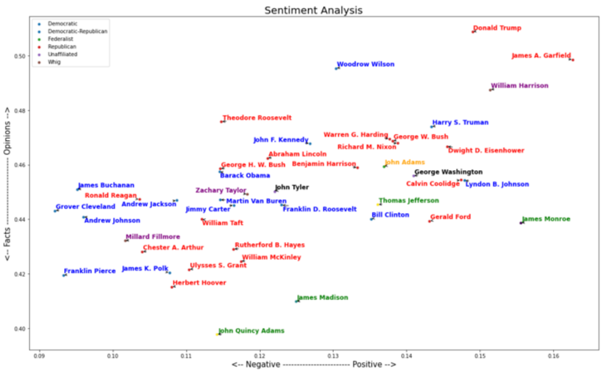
2. PURPOSE STATEMENT AND RESEARCH QUESTION

With this background information and motivation, I want to investigate if a logistic regression model can classify presidential speeches into their correct political party depending on the TF-IDF vectorization and semantic similarity. I think it will be easy for a model to classify speeches into which political party one is because I believe that Presidents of one party tend to speak differently and use different keywords.

3. METHODS

3.1 Data Cleaning

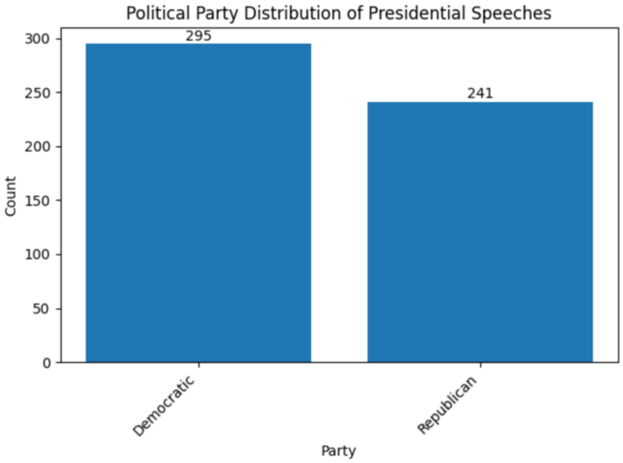
The data used for this assignment was sourced from Kaggle (https://www.kaggle.com/datasets/littleotter/united-states-presidential-speeches?select=presidential_speeches.csv) and offers a comprehensive look at presidential rhetoric. I specifically looked at the presidential_speeches.csv which includes 992 entries and spans 22 MB. This dataset has all of the official presidential speeches in US history up until September 2019. Additionally, Kaggle had some connected notebooks to this dataset that encouraged my research. One user looked at how the sentiment among these speeches and factual vs. opinion-based content changes depending on political parties.



The scatterplot visualizes these findings but shows that political affiliation, denoted by color, does not seem to have a strong uniform correlation with sentiment or fact-opinion balance, indicating that within each political group, there is variation. This is a different take from my own hypothesis that political parties have differing speech patterns.

Back to the dataset, the columns include Date, President, Party, Speech Title, Summary, Transcript, and URL. Since the goal of my project is to correctly classify presidential speeches into 'Democratic' or 'Republican' parties based on the transcript, I needed to clean my data to only include presidents of those two specific parties. I trimmed the data from 992 entries to 867, of those there were 478 Democratic and 389 Republican entries. I also removed the URL and Summary columns as I would like to investigate the raw transcript only. It is also important to see if there are any missing values throughout the data. Luckily, this dataset had information for every column and there were no NA values.

It is also crucial to understand the history and formation of each political party for this assignment. The Democratic and Republican parties practically switched their platforms and ideologies up until the 1920s where they solidified their stances. Natalie Wolchover explains this transition in her article "When did Democrats and Republicans Switch Platforms?" She notes that "initially, in the 1860s, the Republicans, dominant in the northern states, supported expansive federal initiatives like the transcontinental railroad and national currency, aligning with big government principles. Conversely, the Democrats, prevalent in the South, opposed such expansions, advocating for limited federal influence. This alignment shifted notably between Abraham Lincoln's presidency and Franklin D. Roosevelt's tenure." This source helps explain the complex evolution of political identities within the context of U.S. history, emphasizing that the changes were gradual and influenced by socio-economic factors and strategic political positioning. Ever since the 1920s, the parties have stayed consistent in their beliefs with Republicans favoring a limited government approach while Democrats tend to back more robust government interventions and regulations. With all this said, it was vital for me to look at presidential speeches that took place after 1920. I trimmed my dataset to this point and found that there are 536 speeches. Of those 536, 295 were given by Democratic presidents and 241 by Republicans which created a pretty equal sample.



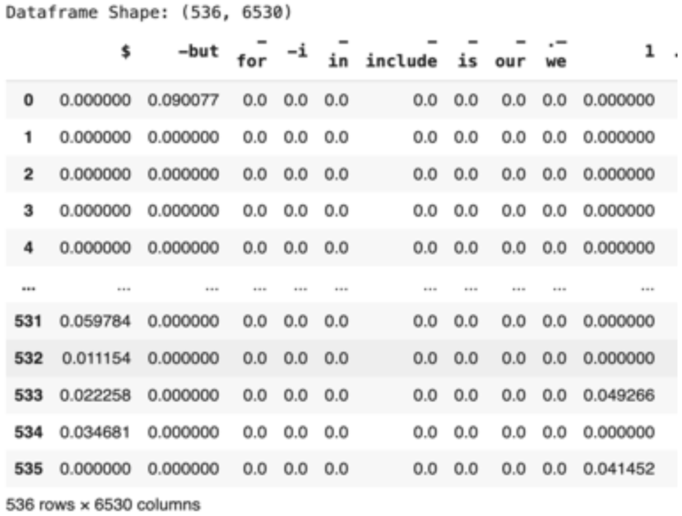
I took some initial counts of this sample and found that the mean sentence count for all these speeches was 172.22 and the mean word count was 3,393.51 while the standard deviation was 2,492.32.

3.2 TF-IDF Analysis

To get started on my analysis, I needed to process the transcript of each speech. I did this by loading Spacy into and a small English language model from the spaCy library. The en_core_web_sm is one of the pre-trained model packages provided by spaCy, which includes a collection of linguistic annotations and capabilities (like part-of-speech tagging, named entity recognition, etc.). I then fed in every speech to the function nlp.pipe() which will process the transcript and create a new column called 'Docs' in my data frame. Each Doc contains a sequence of token objects that have additional linguistic annotations.

Nlp.pipe() has strong capabilities and can process text quicker because it batches them. That being said, trying to feed in around 500 speeches that have on average 175 sentences took a very long time and used a lot of resources. My code continued to break because I was using all the available RAM. To combat this problem, I truncated every speech down to the first 3,500 words. This was the max number of words that I could feed into the nlp.pipe function for all speeches that would run successfully. My cleaned data frame then contained eight columns: date, president, party, speech title, decade, transcript, truncated transcript, and docs.

Following the preprocessing of the U.S. Presidential speeches through the spaCy pipeline, the next phase of my analysis involved converting the processed documents into a structured numerical format that would be amenable to machine learning techniques. To this end, I employed the bag-of-words (BoW) model using the CountVectorizer from scikit-learn, tailored to tokenize documents based on the previously lemmatized and lowercased tokens while excluding stopwords and infrequent words. This resulted in a sparse matrix where rows correspond to documents and columns represent token frequency, effectively capturing the lexical variety within the corpus. In the image below, you can see that there are 536 rows (for each speech) and 6,530 columns which indicate every possible token. The values of each cell show how often each token is in every speech.



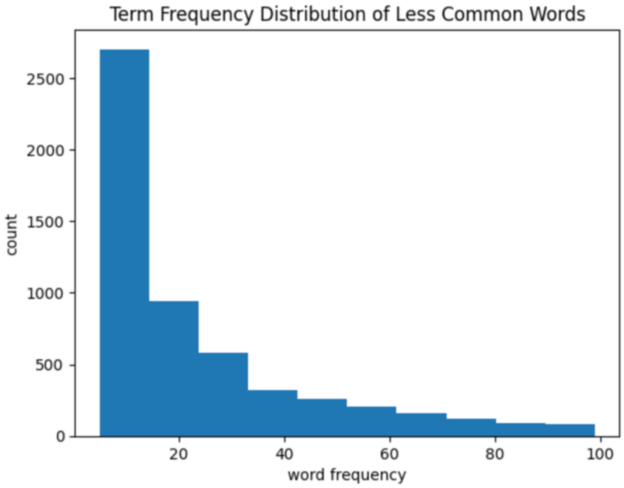
Dataframe Shape: (536, 6530)

	$	-but	—for	-i	—in	—include	—is	—our	.—we	1
0	0.000000	0.090077	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
1	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
2	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
3	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
4	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
...	...	...	...	...	...	...	...	...	...	...
531	0.059784	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
532	0.011154	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
533	0.022258	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.049266
534	0.034681	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
535	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.041452

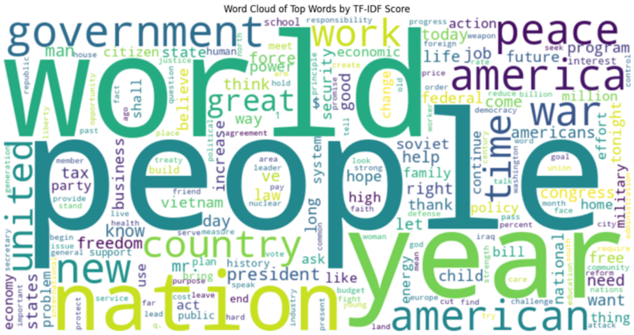
536 rows × 6530 columns

To better understand the distribution of terms across documents, I visualized the term frequencies, revealing a Zipfian distribution—a few words are very common, while most are rare, a pattern typical of natural language.

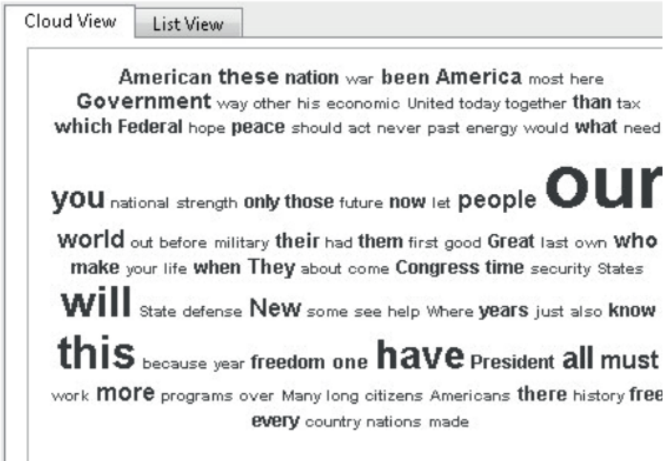
The distribution below shows that a lot of words occur a couple of times, but fewer words show up more than 50 times.



I created an additional visual, a word cloud, that shows the most common 100 words among all the speeches (both Democratic and Republican).



If you were to compare this word cloud to the one from Randy Ramirez's research findings, you can see that there is a difference in common words. Ramirez's findings show that 'our' is the most popular word in 18 different speeches, which 9 were from Republican and 9 from Democratic presidents. There are many words that are found in both word clouds but the most popular ones are very different. This could be because I have many more speeches sampled than Ramirez and that I truncated my speeches to the first 3,500 words but it is interesting to compare the visuals.


Cloud View
List View
American these nation war been America most here
Government way other his economic United today together than tax
which Federal hope peace should act never past energy would what need
you national strength only those future now let people our
world out before military their had them first good Great last own who
make your life when They about come Congress time security States
will State defense New some see help Where years just also know
this because year freedom one have President all must
work more programs over Many long citizens Americans there history free
every country nations made

Further refining the BoW model, I applied the Term Frequency-Inverse Document Frequency (TF-IDF) transformation using TfidfVectorizer. TF-IDF adjusts counts based on how unique a word is to a document in the context of the entire dataset, thus emphasizing words that are distinct to certain documents and potentially more reflective of their content.

Equipped with a rich, numerical representation of the speeches, I proceeded with a modeling phase. To predict the political party of the speaker based on their speech content, I used a logistic regression classifier. This model choice is informed by its simplicity and interpretability, which are valuable this type of research. Using a pipeline setup, which streamlines the vectorization and classification process, I trained the model on a subset of the data for an 80-20 split, ensuring a balance between training and test sets.

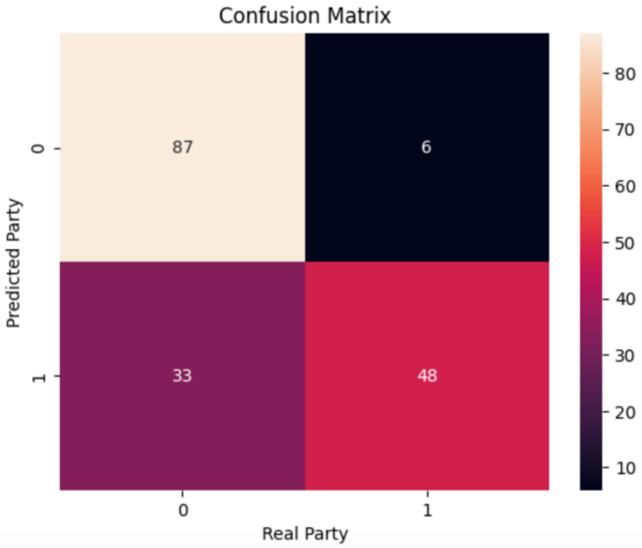
The summary of the findings from the classification report and confusion matrix indicates that the model was more precise in identifying Republican speeches (0.82) than Democratic ones (0.71), but had a better recall for Democratic speeches (0.90). The F1-scores, which combine precision and recall, are fairly balanced across both parties, with Democrats at 0.79 and Republicans at 0.66. The overall accuracy of the model is 0.74.

```
              precision    recall  f1-score   support

  Democratic       0.72      0.94      0.82        93
  Republican       0.89      0.59      0.71        81

    accuracy                           0.78       174
   macro avg       0.81      0.76      0.76       174
weighted avg       0.80      0.78      0.77       174
```

From the confusion matrix, we can infer that the model was more likely to mistake Republican speeches for Democratic ones than vice versa.

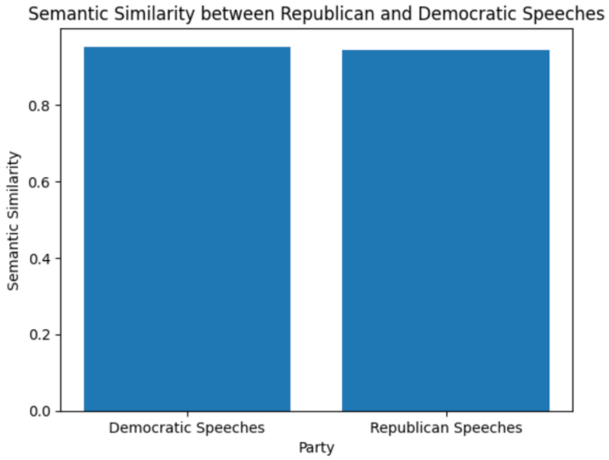


Using TF-IDF to classify presidential speeches has been moderately successful, and these findings could suggest that the model can distinguish between the two parties' speeches based on content, although there is room for improvement, particularly in terms of reducing misclassification of Republican speeches.

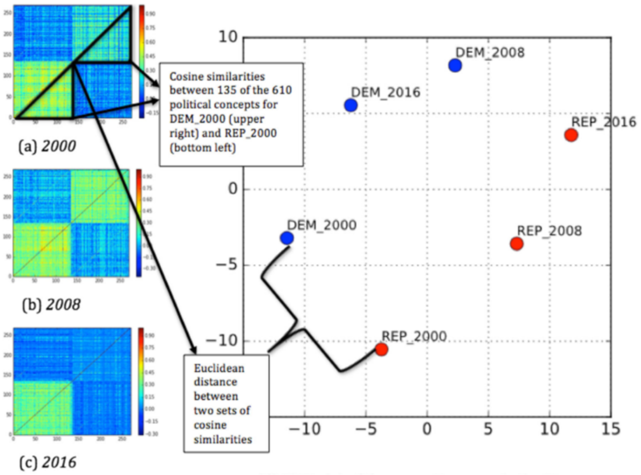
3.3 Semantic Similarity Analysis

In exploring the linguistic landscape of presidential rhetoric, I assessed the semantic similarity within and between the speeches of Democratic and Republican presidents. The objective was to understand if the content of speeches from presidents within the same party shows a greater similarity compared to those from different parties, considering not only the vocabulary used but also the contextual meaning behind the words.

Using the spaCy library, I converted the speeches into numerical vectors that encapsulate their semantic content. I then analyzed semantic similarity among 295 Democratic and 241 Republican speeches by comparing each speech with the next in sequence within the same party group. The analysis revealed substantial semantic coherence within each party: the average similarity score was 0.9518 for Democratic speeches and 0.9435 for Republican speeches.



This suggests a strong consistency in thematic and stylistic elements among presidents of the same party, echoing findings from Li, Schloss, and Follmer's study, which identified "distinct, systematic word association patterns" within political groups that could be "reliably distinguished using machine learning methods" (Li et al., 2017). The diagram below shows their findings. The MDS plot demonstrates that the semantic spaces of the two political parties (DEM for Democratic and REP for Republican) can shift significantly from one election to another. For example, "DEM_2008" and "REP_2008" are closer to each other than "DEM_2000" is to "REP_2000", suggesting that in 2008 the Democrats and Republicans had more in common, semantically, than they did in 2000.



To determine if these intra-party similarities were matched by inter-party differences, I conducted a two-sample t-test on the similarity scores between the parties. The results yielded a t-statistic of 2.498 and a p-value of 0.0129, confirming significant differences in semantic similarity between Democratic and Republican presidential speeches. This significant distinction aligns with Li et al.'s observations that even within a common political discourse, "presidential nominees had distinct conceptual representations" that did not always align with their party's typical rhetoric (Li et al., 2017). The implications are clear: while presidents from the same party share linguistic ties, the semantic gap between parties underscores a deeper ideological divergence. These findings not only support the hypothesis that political affiliation influences presidential speech content, but also suggest that semantic analysis, as a tool, can effectively reveal underlying patterns in political discourse, affirming the pivotal role of language in mirroring and shaping political ideologies. This analysis provides a nuanced understanding of how language reflects the evolving dynamics within and between the major political parties in the U.S.

4. CONCLUSION AND FUTURE WORK

This study embarked on a multifaceted exploration of presidential rhetoric, examining the extent to which linguistic patterns in presidential speeches can be classified according to political party lines. Employing TF-IDF vectorization and semantic similarity analyses, the investigation revealed that logistic regression models are indeed capable of discerning the political party from speech content with moderate success. The linguistic nuances that differentiate party affiliations were underscored by the notable semantic coherence observed within each party's corpus of speeches, with Democratic addresses yielding an average similarity score of 0.9518 and Republican speeches presenting a comparable consistency with a score of 0.9435. Statistical analysis lent further credence to these findings. A two sample t-test comparing intra-party similarities produced a significant t-statistic of 2.498 and a p-value of 0.0129, thereby rejecting the null hypothesis and confirming a discernible ideological divergence in the semantic content of speeches from Democratic and Republican presidents. This distinction aligns with Li et al.'s (2017) assertion that "presidential nominees had distinct conceptual representations," which reinforces the notion that political rhetoric is indeed infused with party-specific ideologies.

Looking ahead, the study aims to extend the current analysis to understand how historical contexts and specific time frames influence presidential discourse. An investigation into speeches surrounding pivotal events — such as the civil rights movement, the Cold War, and major economic shifts — could offer deeper insights into the interplay between political action and presidential rhetoric. This approach will endeavor to decipher how moments of national significance are framed by presidents from different eras and affiliations, potentially revealing how leaders linguistically navigate the waters of public sentiment and policymaking during times of societal change.

In sum, by integrating computational linguistics with historical analysis, this research underscores the profound impact of presidential speech-making on American society. The findings not only shed light on the political stratification of presidential communication but also pave the way for future research that might employ predictive analytics to further unravel the complex tapestry of political language and its implications for leadership and public engagement.

5. REFERENCES

[1] Beasley, Vanessa. *"*Words Matter: What an Inaugural Address Means Now" *The Vanderbilt Project on Unity & American Democracy.* January 15, 2021. https://www.vanderbilt.edu/unity/2021/01/15/words-matter/

[2] ChatGpt4 helped me in summarizing literary references and code

[3] Dawley, Hunter, "How Do Presidents Speak? An Analytical Approach to Presidential Rhetoric." Vanderbilt University, https://github.com/hunterd3/presidential-speech-analysis

[4] Kilber, Tara, "The 15 Most Inspiring Presidential Speeches in American History," *Hein Online*. February 15, 2021. https://home.heinonline.org/blog/2021/02/the-15-most-inspiring-presidential-speeches-in-american-history

[5] Li, P., Schloss, B. & Follmer, D.J. Speaking two "Languages" in America: A semantic space analysis of how presidential candidates and their supporters represent abstract political concepts differently. *Behav* Res 49, 1668–1685 (2017). https://doi.org/10.3758/s13428-017-0931-5

[6] Pennebaker, James W., et al. "Examining long-term trends in politics and culture through language of political leaders and cultural institutions." *Proceedings of the National Academy of Sciences,* February 11, 2019, www.pnas.org/doi/full/10.1073/pnas.1811987116

[7] Ramirez, Randy B. *Protest American Leadership Word Patterns in Presidential Speeches from 1945 Through 2016: A Historical Case Study*. University of Phoenix, 2019. ProQuest Dissertations Publishing, www.proquest.com/docview/2491700181?pq-origsite=gscholar&fromopenview=true&sourcetype=Dissertations%20&%20Theses

[8] "State of the Union: The Words Used." *The New York Times*, The New York Times Company, 25 Jan. 2011, archive.nytimes.com/www.nytimes.com/interactive/2011/01/25/us/politics/state-of-the-union-words-used.html

[9] "25 Decade-Defining Events in U.S. History." *Encyclopedia Britannica*, Encyclopedia Britannica, Inc., https://www.britannica.com/story/25-decade-defining-events-in-us-history