SENTIMENT OF THE UNION

by
Chase Rydeen

A thesis submitted to the faculty of The University of Mississippi in partial fulfillment of
the requirements of the Sally McDonnell Barksdale Honors College.

Oxford
May 2018

Approved by

_____
Advisor: Dr. Dawn Wilkins

_____
Reader: Dr. Naeemul Hassan

_____
Reader: Dr. Yixin Chen

ABSTRACT

As the machine learning and data science craze sweeps the nation, the implications and implementations are vast. This paper takes a look at both of them through the lens of a topic of national importance, at the very least for the United States. This topic is the words used by past Presidents of the United States, which are being pulled from their State of the Union Addresses. These two broad subjects are implemented in varying degree by means of Natural Language Processing (NLP), of which this paper is centered around. Natural Language Processing pulls heavily from both of these two categories to enable effective analysis of text-based data. Using NLP, a sentiment analysis was conducted on the Addresses to gain further insight into the tone used by Presidents over the course of history. This paper shares the methodology used to conduct this sentiment analysis and discusses how it was presented and visualized.

DEDICATION

To be completed.

ACKNOWLEDGEMENTS

Thank you so much to Dr. Wilkins for guiding me on this long journey and always being there when I need her. You are a true inspiration and have been so influential in my undergraduate career, and I'm so thankful for all your advice and help with this paper and this thesis.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Language is our main way of communication and the words an individual chooses to express their ideas can be very telling about their mood and opinion towards a particular subject or issue. This is extremely pertinent for the President of the United States as their words and messages set the tone for the United States as a whole, and their words can often be analyzed to look for further meaning in the things that they say. This analysis has the potential to reveal underlying themes and patterns of speech in how president's speak and what their word choice indicates about the State of the United States and also how they view certain issues and topics. The driving force behind the research conducted here was to see if there was a way to reliably predict a president's political party purely based on their words they use in their State of the Union Addresses. This research can extend far beyond this central question as well, expanding to include more data sources to increase the accuracy of the predictions. On top of this, there is potential to predict further characteristics beyond political party such as ideology and other personality traits of the speaker.

The other crucial part of this research is more historical in nature in that each sentiment score derived for a president must be contextualized by the important historical events that occurred during their presidency. The biggest challenge here being separating out the relative importance of the president's own outlook on the world versus that of the event itself. Some presidents may have a more positive tone and outlook on the world and try to use their State of the Union address to encourage the public even if the events of the time may be dire such as the Great Depression, so that will be an interesting challenge to weigh the relative importance of each.

CHAPTER 2

CORPUS

2.1 State of the Union Addresses

The main textual data that was collected to be processed was all of the Presidential State of the Unions from George Washington's first address to Barack Obama's last address. The text source was initially pulled from a Presidential Address Repository. The text came in a large text file that contained every speech and it was split in to individual text files for each individual address to allow for easier processing. The addresses vary widely in length and content, which is also of significant note when analyzing and comparing these addresses across the timespan of the existence of the United States. George Washington's first address was just over a thousand words and seventeen paragraphs, whereas Barack Obama's final address was just over 5,400 words and was 78 paragraphs long.

2.1.1 Change in Purpose of State of the Union Address

The length is the most notable change in the State of the Union Addresses over time, but there are important factors to consider as well that could potentially impact how the addresses are given from year to year. When the Presidential Addresses first started with George Washington, it was not supposed to be a recurring event [TETEN (2003)], but it soon began to grow to a tradition so that the president could publicly address the people and inform them of the current events of the country. Over the years, the Presidential Address has taken on many forms, in spoken word, in written letter, in radio broadcast, and, nowadays, on live television broadcast. The Presidential Address shifted from the yearly Presidential update, sometimes the only times people would hear directly from the President,

to a formalized briefing to inform the public in an organized manner of the current state of affairs and push forward a President's agenda for the upcoming term [TETEN (2003)]. While that has always been somewhat of the purpose of the addresses, it has become more of the central focus over the course of time due to technological innovations and changes in media coverage. Nowadays, citizens of the United States can read in real-time about the decisions of the Presidency and the Presidents political moves without needing to listen to a one time speech to become updated on their agenda and goals for the year to come. It is a subtle, yet interesting shift in how the addresses are approached and given, but even these purposes could change between years depending on the person giving these orations, an important factor to consider also.

2.1.2 Presidential Personality

Another important factor in how the Presidential State of the Union Addresses are given is the personality of the President that is giving them. This is a rather intangible element of the speeches that can be hard to quantify but is very important to note. Most people have a certain disposition towards being on the more optimistic or pessimistic side of life and that can display itself in the speeches given. Since the important topic being considered here is tone, that can be heavily influenced by if the President giving the speech tends to be more realistic or optimistic in their outlook on the world, and to potentially create a better one. Some President's may see the State of the Union as a chance to rally the nation and drum positivity and support for their platform for years to come, whereas others might see it as a good opportunity to have a nation-wide reality check and bring the citizens in-line with what needs to be done for the good of the nation [TETEN (2003)].

2.2 Statistical Summary

It is important to have an understanding of the speech data itself before diving in to this research, since otherwise it won't be as meaningful and will be harder to draw conclusions from it. The full statistical summary for the data can be found in Table 2.1. This can be

explored to search for trends in the data and familiarize oneself with an overall perspective on the data. Some information of note: There are a total of 230 Presidential Addresses given by 42 Presidents, making the average number of addresses per president 5. There are 26 Republicans and 16 Democrats , which makes their percentages 62% and 38%, respectively. The first three columns are self-explanatory but the latter two may require some additional information to be understood correctly.

2.2.1 Lexical Diversity

Lexical diversity is a metric that is used to represent the amount of unique words in any given passage of text. Lexical diversity is calculated by dividing the number of unique words in a text by the total length of the text. The resulting number is between 0 and 1 and the closer to 1 the text is, the more diverse its lexicon is, and thus it can be interpreted as being more complicated to read. The patterns here can be confounding by sheer length of a text but it remains an important metric to see how complex a particular selection of text is. The nature of this calculation makes it more interesting when comparing two pieces of text that are similar in length to see difference in lexicon between the two.

2.2.2 Grade Level

Calculating grade level is a slightly more involved process that involves an algorithm that computes grade level based on two factors: average sentence length and average syllables per word. The first factor is relatively easy to calculate, but the second is slightly more tricky as syllables can be a lot more difficult to distinguish in plain text processing fashion. Luckily, there is a Python plugin called textstat with a built-in Flesch-Kincaid function that has a corpus of syllabled words and it was used to calculate grade level. You can see how the formula is used in Equation 2.1.

$$0.39 \; (\frac{total \; words}{total \; sentences}) + 11.8 \; (\frac{total \; syllables}{total \; words}) - 15.59 \qquad (2.1)$$

2.3   Information Visualization

An important part of this research is also concerned with how best to display the resulting information in an effective and easy-to-understand manner. There is an entire field dedicated to how to best display technical information and data and how to convey it to large groups of people with little technical background [Fekete et al. (2008)]. This is important with data such as the sentiment score being processed here, as the long numbered sentiment scores are intimidating and without any context, data is meaningless. The context here is contained within the graph used to display the sentiment score data and interactive features were implemented to help users engage with the data in a more meaningful fashion. The data in this research is quantitative and since the Presidential Addresses are given in chronological order, time was used on the x-axis and the data lended itself nicely to a Line Plot. This line plot will be discussed more in-depth in the following section.

2.3.1   Word Cloud

A word cloud is a collage of words that displays word frequencies for a certain set of text data, with the relative size of each word being determined by the frequency with which that term is used in the text. An example can be seen in Figure 2.1 that shows the word cloud for all of Jimmy Carter's terms. A second example can be seen in Figure 2.2 where a term is selected and the word cloud dataset is restricted to the contents of that particular presidential address. Word Clouds are an interesting visual since they provide quick reference to see what a President's most used terms are, as well as being another interesting way to engage the data in a slightly different context. Word Clouds themselves are often criticized since it is a poor way to visualize data and it is hard to objectively compare two words in a word cloud because the frequency values are encoded using area, which is a very difficult encoding for humans to interpret. In this case, the word cloud is used merely to complement the existing line plot visualization that provides insight into the main purpose of the research, and the word cloud provides a different way of visualizing the data source itself.

5

President	# of Addrs.	Avg # Words	Lex Diversity	Grade Level
George Washington	8	2096.0	0.3762	18.55
John Adams	4	1801.0	0.369	17.925
Thomas Jefferson	8	2605.0	0.3376	18.0
James Madison	8	2729.0	0.3433	20.825
James Monroe	8	5326.0	0.2493	16.462
John Quincy Adams	4	7864.0	0.2327	19.25
Andrew Jackson	8	10708.0	0.2042	19.2
Martin van Buren	4	11411.0	0.2036	20.15
John Tyler	4	8560.0	0.2291	18.475
James Polk	4	18173.0	0.1525	17.275
Zachary Taylor	1	7678.0	0.2346	17.2
Millard Fillmore	3	10612.0	0.2224	16.967
Franklin Pierce	4	10545.0	0.2192	19.15
James Buchanan	4	14247.0	0.1797	15.05
Abraham Lincoln	4	6999.0	0.2639	13.675
Andrew Johnson	4	9690.0	0.2294	15.9
Ulysses S. Grant	8	8232.0	0.2391	15.938
Rutherford B. Hayes	4	8692.0	0.2363	16.325
Chester A. Arthur	4	5045.0	0.3252	13.6
Grover Cleveland	4	12478.0	0.2236	17.45
Benjamin Harrison	4	13881.0	0.1976	14.7
Grover Cleveland	4	14969.0	0.2121	16.35
William McKinley	4	16901.0	0.1977	15.8
Theodore Roosevelt	8	19793.0	0.1732	14.975
William H. Taft	4	17594.0	0.1868	17.025
Woodrow Wilson	8	4384.0	0.2768	15.05
Warren Harding	2	5738.0	0.2768	13.5
Calvin Coolidge	6	8707.0	0.2306	11.783
Herbert Hoover	4	6489.0	0.2566	14.15
Franklin D. Roosevelt	12	3991.0	0.3002	12.0
Harry S. Truman	8	8405.0	0.2321	10.475
Dwight D. Eisenhower	9	6103.0	0.2751	12.3
John F. Kennedy	3	5816.0	0.289	12.233
Lyndon B. Johnson	6	4917.0	0.2707	10.017
Richard Nixon	5	4002.0	0.2692	11.78
Gerald R. Ford	3	4649.0	0.2865	10.767
Jimmy Carter	4	11410.0	0.2427	11.05
Ronald Reagan	7	4731.0	0.2963	9.557
George H.W. Bush	4	4396.0	0.285	7.8
Bill Clinton	8	7528.0	0.2207	9.35
George W. Bush	9	4888.0	0.2883	9.122
Barack Obama	8	6738.0	0.2465	8.412
Donald Trump	1	788	6344	0

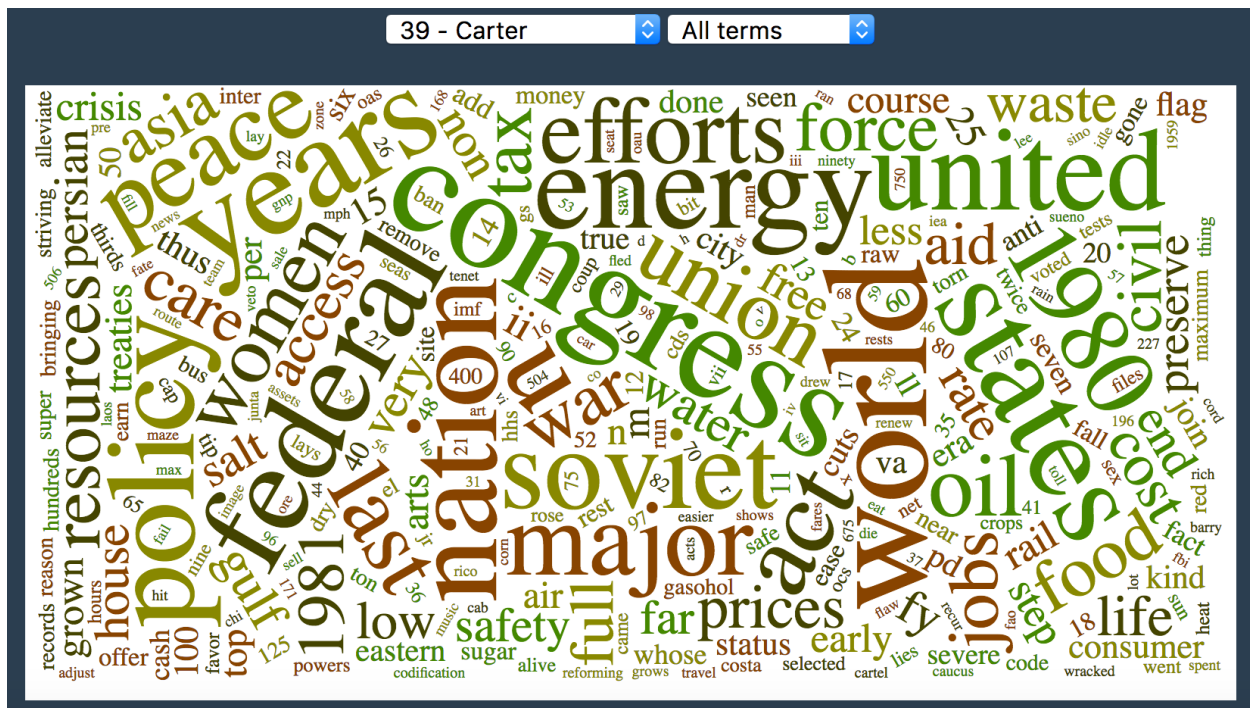Table 2.1: Presidential Summary Statistics

Figure 2.1: Example Word Cloud showing all terms for Jimmy Carter

Figure 2.2: Example Word Cloud showing 1978 term for Jimmy Carter

2.3.2 D3

D3.js (D3) is the JavaScript Library used to create the visualization mentioned in the previous section and another mentioned in a future section. D3 uses pre-built JavaScript functions to select, elements, create SVG elements, style them, or add dynamic effects or tooltips to them [Bostock et al. (2011)]. D3 also has a handy library that creates word clouds that was used in this research, it takes in an input array in JSON format with the words and their frequencies in decreasing order and draws the words with their relative sizes on the HTML canvas. There is a bit of a delay on the drawing of the word clouds, since instead of saved images of the word clouds, the program is actually drawing all of them in realtime and just swapping out the JSON data source depending on which President is selected in the dropdown menu at the top of the page.

2.4 Results

The bulk of this information is used later for processing, but it is important to understand the data source as well, before diving in to predictions using it. The importance and purpose of the State of the Union address is important here since it has a heavy-handed influence on the content and message behind the Presidential Addresses. An important note to make is of the party breakdown, with 26 Republicans and 16 Democrats , which makes their percentages 62% and 38%, respectively. These numbers aren't exactly correct, as the lesser-known and ephemeral early parties were placed in to either Democrat or Republican based on their policy positions. This was done in order to maximize the effectiveness of the prediction algorithm that will be discussed later on. The statistics here are important as they provide more insight into the data being processed and provide a more concise view into what is being handled. Table 2.1 has some intriguing patterns and trends to analyze and show.

2.4.1 Average Address Length

The average length of the Presidential Address has changed drastically over time, as its purpose and importance fluctuated. George Washington, when he gave his first address, didn't think that it would be an occurring event, but thought it necessary to inform the citizens of the current state of affairs of the country, and this precedent was followed for much of the early history of the United States. The relatively short length of the early Presidential Addresses shows this, as it was a short and brief and meant to inform the people of what is happening in the country and was used primarily to disperse information to the citizens. This slowly began to change over the years and the change in purely the average number of words in the addresses indicates the increasing importance of the State of the Union address as a chance to communicate with the citizens at large and use that attention to push an agenda and connect with the voters. The State of the Union address became a much larger deal as President's used it to communicate with the entirety of the nation to ensure them of the success of the nation and its status, peaking with Theodore Roosevelt averaging almost 20,000 words per Presidential Address. Shortly after, however, the length of the addresses had a tremendous drop-off from Taft at 17,000 words to Woodrow Wilson at 4,300 words, which can likely be attributed to the emergence of World War I. The country was involved in a major war effort and the fanfare and policy pushing of State of the Unions past were cleared out of the way for the focused messages of State of the Unions to come. These were defining times in the world, and with a major conflict to unite all people in the country, the State of the Unions became more condensed and focused on the important aspects on hand to encourage the country and assure them of the success of the war effort and keep country-wide morale high and trying not to distract them for too long. From this major change and in to the modern era, the State of the Union address has stabilized around 5,000 to 10,000 words, keeping to an average length and the TV equivalent of roughly an hour to an hour and a half, long enough to keep people's attentions and effectively convey a president's reflections on the past year and goals for the next.

2.4.2 Lexical Diversity

Lexical Diversity is also interesting to note here and it generally follows the same pattern as average length, just in the reverse fashion. As one would expect, the more words that are spoken, the less overall unique words are going to be spoken. This is most evident when examining the lexical diversity of George Washington and that of Teddy Roosevelt. George Washington had notoriously short State of the Union Addresses so his average lexical diversity was 0.3762, whereas Teddy Roosevelt has an average lexical diversity of 0.1732, which makes sense since his average length is almost ten times greater than that of George Washington's. This provides more important insight in addresses that are similar length to one another and provides a deeper insight into the speech-writing process and how word selection is important when communicating information to large swathes of people and needing to be considerate of their education levels.

2.4.3 Grade Level

Another metric that complements Lexical Diversity that needs to be considered is Grade Level, which is computed using the Flesch Kincaid mentioned above. The scores seen here may seem rather high but it is understandable given the change in how Americans speak over time. Speakers in Early America were known to have a rather complicated way of talking and in order to make it to the office of President one had to be sufficiently educated to get elected by the public. This pattern shows in the high grade level throughout the early and mid history of the United States as a sophisticated approach was often taken to the State of the Union address to show off the strength and power of the United States [CITE]. Also the early speeches were often given in front of Congress and with no means of distributing the speech widely, the intended audience mainly stayed Congress so the early Presidents did not really have a need to simplify their language to communicate effectively to the common man as that was done by the newspapers that talked about the Address [CITE]. The grade level gradually has decreased over time, which has as much to do about

the greater reaches the address has, as it does with modern day media and how it collects sound bites of presidential addresses. In the modern era, when a President gives a speech, only a small amount of the actual speech is rebroadcasting when the media is discussing it, so the "sound bite" phenomenon has arisen in State of the Union addresses, which has had a transitive effect on the Flesch-Kincaid grade level calculation. The media only takes small snippets of what the President says to convey major policy positions, which has had the effect that most statements are kept short in order to summarize points and clearly convey what positions the President has in as short a form as possible. And since one of the calculations for the grade level calculation is average words per sentence, this brings down the grade level of the speech as the President attempts to become more clear in their purpose and position to effectively convey their thoughts and feelings in a short sound bite that could be taken from their speech.

CHAPTER 3

SENTIMENT ANALYSIS

3.1 Sentiment Analysis

The cornerstone for this thesis is sentiment analysis and this topic requires the definition of a lot of terms and concepts that are important to the research at hand. Sentiment Analysis in theory sounds rather simple, process text and pull out the meaning based on the content of what was processed, but there are many intricacies that need to be addressed to fully understand the entire process [Liu (2012)].

3.1.1 Definition

Sentiment Analysis refers to the use of natural language processing and text analysis to systematically identify, extract, quantify, and study affective states and subjective information [Liu (2012)]. Sentiment Analysis has increased in popularity in recent years and is popular to use to review large sets of review / survey data to abstract major topics of conversation and controversy online. It can be an effective tool in summarizing a population's opinions and feelings towards certain issues and drawing conclusions from them. A basic task of sentiment analysis that can be leveraged into more complex tasks is determining the polarity of a sample of text data and classify it as positive, negative, or neutral [Wilson et al. (2005b)]. The process behind sentiment analysis is important and can be complicated depending on how much information an analysis is trying to pull and how large of a dataset is involved, and before the process is addressed, there are certain principles and topics involved that need to be covered first.

3.1.2 Natural Language Processing

Natural Language Processing (NLP) is an important concept that is used heavily in sentiment analysis. NLP is primarily concerned with the interactions between human beings and computers and specifically how they process and discern the meaning of human language [Liddy (2001)]. Natural Language Data is abundant in our world today in the Age of the Internet and the vastness of it makes NLP extremely important to implement effectively to aid in understanding this large dataset. NLP takes on the difficult task of processing large text data and attempting to quantify the text data in different ways.

3.1.2.1 Natural Language Toolkit

NLP is the concept and the implementation in this project is the Natural Language Toolkit (NLTK), which is a Python library that offers NLP methods to process text and extract meaningful trends and patterns from the text of interest. The NLTK is implemented using Python, which is a simple, yet powerful language with excellent functionality for processing linguistic data Bird and Loper (2004). Much of the meaning is derived using the NLTK but much of the processing is conducted in purely Python using lists of words to process meaning and sentiment. The task of processing text data comes with a few obstacles that can either obstruct meaning or complicate the processing by changing the meaning of words or phrases based on the context in which they reside.

3.1.3 Complications with Text Data

Text Data can be especially difficult to deal with and cause a lot of unforeseen issues when it is being processed. The inherent subjectivity of human language and speech is one of the largest obstacles that must be addressed when dealing with any text-based data. Much of the communication between human beings is subjective and it is up to the interpretation of the speaker and listener what the message is and they can have conflicting ideas on the meaning of some terms [Aggarwal and Zhai (2012)]. This potential for miscommunication is mirrored in NLP in that the determined meaning or value of some textual data could not

be representative of the source it came from, and there is not an objective reference to the weight of words to confirm the correctness of any one interpretation of the textual data.

Another important factor to consider when processing text data is sarcasm, which is extremely difficult to detect. Americans especially are known for their use of sarcasm and it can be sometimes impossible to parse such meaning out of text since the intonation is what indicates the sarcasm which is lost in purely text-based data. There are some subtle cues that can indicate sarcasm in text but it can really only consistently be caught if it is tagged as such [Riloff et al. (2013)].

A third and final complication that often arises is considering the context in which the text resides [Aggarwal and Zhai (2012)]. Context is everything when examining how people speak and trying to accurately access the thoughts and opinions of the speaker so it is important to take this context into consideration when developing a sentiment analysis since it will affect any results that are achieved. Context here meaning the modifying words surrounding the word to be examined next in the analysis. Each word in a sample of text data can have any number of modifiers that can manipulate its meaning and fundamentally change what message it is conveying by adding certain words before or after the word. These modifiers can take on the form of intensifiers, such as very, that amplify the meaning of a word,

3.1.4   Lexicon

There are several ways to conduct sentiment analysis, some of which do not require a lexicon but this research used a lexicon-based approach. An important tool that is necessary in conducting this analysis is a comprehensive lexicon. A lexicon is a database of words and accompanying features associated with each word [Taboada et al. (2011)]. These features associated with each word vary widely in what they say about the word, from part of speech to length to polarity score. The lexicon used in this research consists of a word and an associated sentiment score that is on the scale from -1.0 (negative) to 1.0 (positive), indicating

how positive or negative the word is. These sentiment scores were compiled from several different lexicons and the results were compiled by surveying thousands of individuals and having them score a certain subset of words and combine those ratings into an average score for each word [Somasundaran and Wiebe (2010)].

3.1.5 Alternative Sentiment Analysis Approaches

There are other ways of conducting sentiment analysis without the use of a lexicon that can also be useful for conducting the analysis. The main alternative method is a comparative approach that compares each block of text data to one another and instead of giving them an objective score, ranks them according to a subset of rules that determine their ranking relative to the other samples of text [Wilson et al. (2005a)]. This approach puts much more focus on the context of what is being said and uses context to determine the polarity of language. This is the main approach used by many political science researchers, since it is far easier to compare politician's values relative to one another than use an objective dictionary to determine their stance on an issue. [Laver et al. (2003)]

3.1.6 Process

The process behind sentiment analysis, especially when it is lexicon-based, is rather simple to understand but there are a lot of hidden factors that must be considered. The most important and most influential part of this entire process is the lexicon, which was covered in the previous section. The lexicon is used as the basis for all the sentiment scores that are assigned in the analysis, thus its integrity and accuracy is central to the success of the analysis. The analysis is started by feeding in the text data to the program. In this case, that text data was the Presidential State of the Union Addresses. This text data is split into an array, each index corresponding to an individual word. At this point, the text data is cleaned up, using assorted algorithms to handle punctuation and capitalization and other linguistic features that could complicate the sentiment score assignment. Once the data is cleaned, it is time to start the bulk of the sentiment score operation. Using nested for-loops,

each individual word is compared to a list of "common" words to avoid wasting time and processing power on "the" and other non-notable words, and then each non-common word is compared to the lexicon the score for that individual score is recorded in an overall score variable. As each word is processed, this score variable is either incremented or decremented based on the value associated with said word in the lexicon, and also an overall counter variable is incremented each time, counting each individual word. After all the words have been processed, the total score that has been tabulated is then divided by the counter and that resulting number is the sentiment score for that selection of text. There are a couple of odds and ends that were glossed over that will be covered in the next sections since they were added after the initial algorithm was constructed that slightly influence it's behavior when encountering certain specialty words, and also classifying different topics within the body of text.

Input: All State of the Union Addresses
Output: The sentiment score for each Presidential Address for each category.

1 open all .txt files and store them in lists of special category trigger words
2 **for** *each address in the State of the Union Addresses* **do**
3   format address
4   split address in to sentences
5   **for** *each sentence in the address* **do**
6   add sentence to 'overall' category
7   **if** *sentence contains category trigger word* **then**
8   add sentence to category
9   **end**
10   **for** *each category* **do**
11   append list of sentences for that category to an overall list
12   **end**
13   **for** *each topic in the overall list* **do**
14   **for** *each word in the topic* **do**
15   create word count for each word and store it in a dictionary
16   **if** *previous word negator* **then**
17   increment negator counter for that word by one
18   **end**
19   **if** *previous word intensifier* **then**
20   increment intensifier counter for that word by one
21   **end**
22   **end**
23   **for** *each word in the dictionary* **do**
24   **if** *word is in lexicon* **then**
25   **if** *length of negators[word] != 0* **then**
26   Subtract length from total count for that word
27   **end**
28   **end**
29   **if** *length of intensifiers[word] != 0* **then**
30   Raise length number of scores to the power of 2
31   **end**
32   Calculate the Sentiment Score by multiplying the number of occurences of the term by the score in the lexicon.
33   **end**
34   **end**
35   **end**
36 **end**

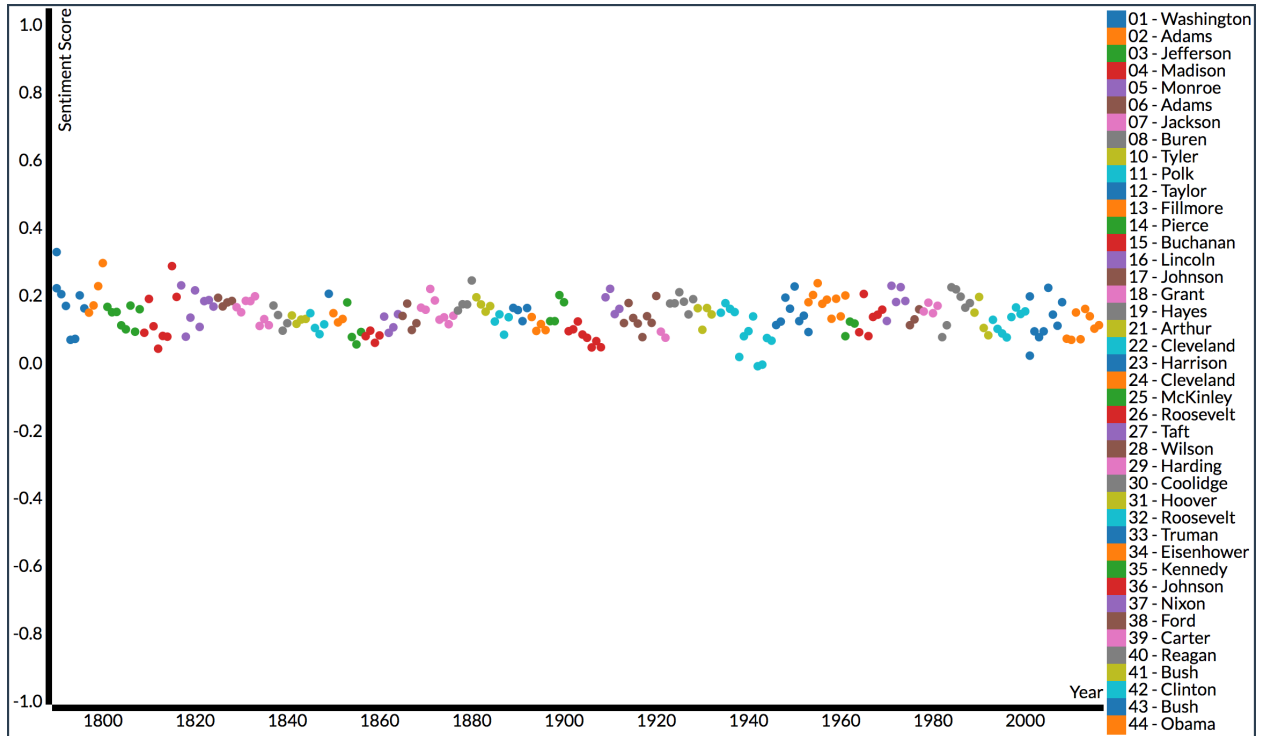**Algorithm 1:** Sentiment Analysis Algorithm

Figure 3.1: Scatter Plot

3.1.7   Scatter Plot

Scatter plots are effective at showing data over time and it allows for users to see overall trends in tone and compare the scores across presidencies to see tonal shifts over a president's tenure or how two presidents compared to one another. The data set lends itself to this representation and the result is a nice longitudinal summary of presidential tones over the course of history. The data being displayed isn't objective and it must be taken with a grain of salt because sentiment analysis is far from an exact science and the lexicon is objective but also doesn't take in to account the change of meaning seen in some words. The time period for these changes is a relatively short period of time in the context of language so the differences shouldn't be greatly significant in the shifting of tone but it is something to note. The scatter plot itself also allows for interaction in that the user can hover over a point and get detailed information about it, such as the President's name, the term and year that address was delivered, as well as the exact sentiment score.

18

3.1.8 D3

D3.js (D3) is the JavaScript Library used to create the two visualizations mentioned in the previous sections. This tooltip functionality was used in the line plot demonstration mentioned previously to give extra information on each data point without cluttering the visualization itself. Another feature implemented using D3 is that if the user hovers over a President's name, then just that President's data points will be highlighted and the rest are faded out of the screen.

3.2 Results

CHAPTER 4

MACHINE LEARNING

4.1   Topic Classifier Sets

An important part of the latter half of the preprocessing work for this research was the topic classifier sets. At first, the sentiment score was calculated for each presidential address with an overall score from -1 to 1, indicating their tone when delivering that address. After these were calculated, they were analyzed to look for trends in each president's tone to see if there were any interesting observations to be seen. As an additional breakdown to see if there was any more context-specific information that could help determine a president's political party, topic categories were added to diversify the scores of the president's even more.

Four Presidential Addresses were chosen (Washington, Lincoln, Kennedy, Obama) and manually read through to discover what words were being used when talking about certain topics within the United States. The topics that were chosen were: crime, economy, education, energy, environment, family, foreign, government, job, religion, terrorism, and war. Text files were created using the trigger words for each major topic, the trigger words being pulled from the four addresses mentioned above and throughout various other addresses as they were skimmed through, that would add the entire sentence to an array named for what topic it was going to collect information on. This processing was conducted on every address and the sentiment score for each topic was found for each president, which resulted in a vector for each address that had their overall sentiment score and the sentiment score for each topic covered in the address. These scores for each address were then averaged together to create an overall vector for each president that could be used for classification

and learning to learn their political party and predict others.

CHAPTER 5

CONCLUSION

5.1 Conclusions

BIBLIOGRAPHY

BIBLIOGRAPHY

Aggarwal, C. C., and C. Zhai (2012), *Mining text data*, Springer Science & Business Media.

Bird, S., and E. Loper (2004), Nltk: The natural language toolkit, doi:10.3115/1219044. 1219075.

Bostock, M., V. Ogievetsky, and J. Heer (2011), D3 data-driven documents, *IEEE Transactions on Visualization and Computer Graphics*, 17(12), 2301–2309, doi:10.1109/TVCG. 2011.185.

Fekete, J.-D., J. J. Van Wijk, J. T. Stasko, and C. North (2008), The value of information visualization, in *Information visualization*, pp. 1–18, Springer.

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

Liddy, E. D. (2001), Natural language processing.

Liu, B. (2012), Sentiment analysis and opinion mining, *Synthesis lectures on human language technologies*, 5(1), 1–167.

Riloff, E., A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang (2013), Sarcasm as contrast between a positive sentiment and negative situation, in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 704–714.

Somasundaran, S., and J. Wiebe (2010), Recognizing stances in ideological on-line debates, pp. 116–124.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, and M. Stede (2011), Lexicon-based methods for sentiment analysis, *Computational linguistics*, 37(2), 267–307.

TETEN, R. L. (2003), Evolution of the modern rhetorical presidency: Presidential presentation and development of the state of the union address, *Presidential Studies Quarterly*, 33(2), 333–346, doi:10.1111/j.1741-5705.2003.tb00033.x.

Wilson, T., J. Wiebe, and P. Hoffmann (2005a), Recognizing contextual polarity in phrase-level sentiment analysis, in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pp. 347–354, Association for Computational Linguistics, Stroudsburg, PA, USA, doi:10.3115/1220575.1220619.

Wilson, T., J. Wiebe, and P. Hoffmann (2005b), Recognizing contextual polarity in phrase-level sentiment analysis, in *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pp. 347–354, Association for Computational Linguistics.

VITA

To be completed.