Contextual Shifts of Political Buzzwords

Studying the Impact of Major Events on the Language of American Presidents

Hope McIntyre
UVA Data Science Institute
hm7zg@virginia.edu

Brian Sachtjen
UVA Data Science Institute
bws7vs@virginia.edu

Nicholas Venuti
UVA Data Science Institute
nmv7de@virginia.edu

ABSTRACT

Words spoken by the president are some of the most impactful in the world, and can influence many outcomes, including facilitating new legislation, changing world policy, and even leading the country to war. Key words, or political buzzwords like "growth" or "security", used within these speeches are critical to how the president communicates with the American people. We sought to research the extent to which semantic density, a quantification of contextual variation, identifies contextual changes in around these buzzwords pre- and post major events. Using the Compilation of Presidential Documents, the team analyzed the presidential rhetoric before and after three major political events: 9/11, the Financial Crash, and the Sandy Hook shooting. Through the context vectors methodology presented by Sagi et al.[4], along with a custom context vector-based clustering technique, we performed an exploratory data analysis to ascertain how well context vectors quantify the magnitude and the nature of the shifts in presidential buzzwords. Using these tools, we were able to identify multiple linguistic trends within presidential discourse during each of the aforementioned events. Furthermore, we were able to confirm the efficacy of both the context vector and the context vector-based cluster technique.

Keywords

Semantic density; context vectors; politics; clustering

1. INTRODUCTION

Words spoken by the president, the most powerful figure in the American political system, are some of the most impactful in the world. The words in his speeches can influence many outcomes, including facilitating new legislation, changing world policy, and even leading the country to war. Key words, or political buzzwords like "growth" or "security", used within these speeches are critical to how the president communicates with the American people. The manner in which these buzzwords are used expresses the president's sentiment and perspective on important topics. While the identification or counting of these words is interesting, there is much more value in extending the analysis to measuring the contextual shifts of these buzzwords, such as what other words tend to be used around them. Understanding the intent and message of political leaders is important both in the political and public policy sphere, as well as in following shifts in societal attitudes. By gaining insight into president's views and the current American state of affairs through this methodology, speeches made by the president can be both retrospective of the American times and potentially predictive of future actions.

With this as motivation, we sought to research the extent to which semantic density, a quantification of contextual variation, identifies contextual changes in presidential discourse pre- and post major events. Our work was based on the work conducted by Sagi et al.[4] which has shown it is possible to measure contextual shifts within language utilizing a method called context vectors. Through our analysis we looked at employing these methods, along with a new context vector-based clustering technique to ascertain how well context vectors quantify the magnitude and the nature of the shifts in presidential buzzwords.

WOODSTOCK '97 El Paso, Texas USA

DOI: 10.475/123_4

2. RELATED WORK

Several text mining approaches have been employed to analyze the content of presidential speeches. These methods include analysis of speech complexity and topical modeling. In a study by Vocative [1], the authors assigned grade reading levels to over 600 presidential speeches and analyzed the temporal changes in these speeches. This was done by applying the Flesch-Kincaid readability test (a measure of text complexity based on word to sentence ratios and syllable to word ratios) to measure the complexity of presidential speeches over time. Through their analysis, the researchers found that the average grade level for presidential speeches has continually decreased over time, with speeches pre-1850 recording at a college reading level, and speeches after 1940 recording at a sixth grade reading level. While this insight is interesting, this broad sweeping approach does not allow for the comparison of semantic meaning between documents.

In a study conducted by Michael Heilman at Civis Analytics [2], log entropy weighted term matrices were constructed from State of the Union addresses to calculate the topical similarities between the presidents' speeches. Through his analysis, he was able to identify ideological and temporal trends within the way the president addressed his constituents during these annual events. For example, he was

able to find that the speeches temporally correlated with one another (i.e. the verbiage used by recent presidents were more similar than the verbiage used by presidents serving further away from each other in time). He was also able to track how a single president's vernacular within State of the Union addresses changed during his term. He also found that President Obama focused mainly on small business development, renewable energy and politics during his first State of the Union Address. He pivoted to student loans, Syria, and terrorism during his final State of the Union address. By focusing on the literal words appearing in these texts, this approach captures broad topical changes between documents, but it fails to indicate how the usage of specific words changed over time.

3. METHOD

To extract the variations in word usage, we developed a representation of the context vector algorithm presented by Sagi et al.[4]. Context vectors allow us to measure semantic density, which is a measure of the range of contexts in which a word appears. This lends to a parallel association with the idea of density; a word with tightly "packed" context vectors in hyperspace indicates a high semantic density. Words with a high semantic density appear in a relatively small number of contexts. Conversely, words with distant context vectors in hyperspace have a low semantic density, which appear in a relatively high number of contexts.

3.1 Pre-processing

In order to develop the context vectors, first we pre-processed the raw text. To do this, we normalized the text by tokenizing, stemming, and finally removing the punctuation and numbers. Stopwords were not removed in this analysis, as doing so would change the spatial relationships between words, and could result in inaccurate results.

3.2 Co-Occurrence Matrix

Next, a co-occurrence matrix was constructed to capture how often each word appeared with other words in the vocabulary. This was done by iterating over each token in the set of documents associated with the event and counting the words within +/- a k sized window of the token. For this analysis, a k value of 6 was used. If the token was near the beginning or end of the document, the window was truncated in that respective direction to ensure that words occurring at the beginning or end of a document were not penalized. The result of this was a hashmap containing the full set of tokens with a count of the times each other token in the vocab occurred within a k-sized window of it. The values stored in this hashmap are co-occurrence vectors, denoted as Vi, for each word i over a vocabulary of length v can be represented as such:

$$V_i = [x_{i1}, x_{x2}, ..., x_{iv}]$$

where x_{iv} is a word in the vocabulary that co-occurs with V_i.

3.3 DSM

Once the co-occurrence matrix was constructed, a distributional semantic matrix (DSM) was developed to reduce the computational load. This was done by converting the co-occurrence hashmap to a sparse matrix. The resulting sparse-matrix was then reduced down to 100 components using truncated SVD. We opted to reduce our feature space to 100 components as it was the same value used in the work by Sagi et al.[4]. Following SVD, the DSM represented as a sparse matrix was converted to a hashmap for quick information retrieval.

3.4 Context Vectors

Following the development of the DSM, context vectors were created. The context vectors for each target word were developed by extracting the words within a k sized window surrounding the target word and summing the DSM hashmaps for the words within the window. As a result, the context ci represents information about the context in which a target term wi appears. For each of the l occurrences of this target term, the context vector is defined as:

$$c_{ij} = \sum_{k \epsilon window(w_i)} dsm(w)$$

3.5 Semantic Density

To calculate the semantic density for each target word, we extracted the context vectors for that term. We then estimated the average cosine similarity of all the context vectors by randomly sampling 2 context vectors from the set and calculating the cosine similarity between them. This was performed over n=1000 iterations. The resulting values were averaged; this was used to represent the semantic density for that target word:

$$score(w) = \sum_{k=1}^{l-1} \frac{< c_{ik}, c_{i(k+1)}}{\|c_{ik}\| * \|c_{i(k+1)}\|}$$

3.6 k-Nearest Context Vectors

For each target word, the k-Nearest context vectors (k-NCV) were calculated. To find comparisons, we followed the method provided by Sagi et al.[4], which involved ranking vocabulary by term frequency and selecting a subset of the most frequent words, known as "content-bearing words." Their motivation was to remove head terms that don't have much discriminatory power (i.e. stopwords), while also removing tail words that only appear in a few documents. For our analysis, the 50th-250th most frequent words were used as this subset. For each of the content bearing words and the target word, the context vectors were summed together and divided by the magnitude to produce a representative unit vector for each word. The cosine similarity was calculated between the target's representative unit vector and the representative unit vector for each word in the set of content bearing words. Using this metric, the k = 5 most similar words associated with the target's representative unit vectors were selected.

4. DATA

The Team used a set of text from the Government Publishing Office [3], the Compilation of Presidential Documents, which includes all public remarks made by the president from January 1992 to March 2016. Documents included press releases, memos, and major speeches like the State of the Union. A text file of each document was downloaded from the website along with metadata, such as the date and title of the remarks.

5. EXPERIMENT

5.1 Time Windows

To validate the idea that political buzzwords change over time, the team analyzed the shifts in semantic density for target words in response to three major historical events of the last 20 years: the terrorist attacks of 9/11, the 2008 financial crisis, and the Sandy Hook mass shooting. These particular events were chosen because they were substantial events for impacting the American public, and also had significant political ramifications. For each one of these events, a time window was selected before and after each event to analyze the shifts associated with the ramifications of these events, as given in the Table 1 below. The Financial Crisis also has a "during" window because it was not a one-time event, but rather a series of cascading incidences over the period of several years.

Table 1: Time Windows

Event	"Pre" Window	During Window	"Post" Window
9/11	9/11/1999 - 9/10/2001		9/11/2001 - 9/11/2003
Financial Crisis	1/1/2005 - 12/31/2006	1/1/2007 - 12/31/2008	1/1/2009 - 12/31/2010
Sandy Hook	9/14/2012 - 12/13/2012		12/14/2012 - 3/14/2013

For the 9/11 and Financial Crisis events, a two year window was chosen for the before, during, and after periods because we felt that two years was a sufficient time span in order to capture contextual usage of our target words. An alternative window size was used for the Sandy Hook analysis in an effort to avoid overlapping with other mass shooting events that occurred. Three months was the maximum window we could use that would satisfy this constraint. After defining our windows, the corpus was subset to only those documents that occurred in that window. We also reduced our subset to only documents that contained at least one of our target buzzwords, in order to reduce our search space.

5.2 Target Words

To create the context vectors, the team first curated a list of topic-specific buzzwords we felt were associated with the specific topic, as shown in Table 2 below. The goal here was to use both general presidential buzzwords such as "defend", as well as event-specific buzzwords such as "Iraq".

Table 2: Buzzword List

Event	Buzzword List
9/11	'terrorism', 'laden', 'qaeda', 'wmd', 'attack', 'homeland','security', 'defend', 'islam', 'freedom', 'iraq', 'afghanistan', 'peace', 'war', 'protect', 'god'
Financial Crisis	'business', 'loan', 'economy', 'bank', 'bailout', 'stability', 'stimulus', 'tax', 'billion', 'mortgage', 'recovery', 'stock', 'street', 'unemployment', 'jobs', 'foreclosure', 'treasury', 'regulation', 'greed', 'recession', 'credit'
Sandy Hook	'gun', 'gunman', 'shoot', 'tragedy', 'background', 'mental', 'hate', 'safety', 'god', 'death', 'kill', 'hatred', 'control', 'congress', 'amendment', 'right'

5.3 Semantic Density

Using the method described in the sections above, semantic density was calculated for each buzzword in their representative time window in order to identify the magnitude and direction of the shifts in context. Reductions in cosine similarity across time windows would indicate a decrease in semantic density (i.e. the buzzword began being used in more contexts) and an increase in cosine similarity would indicate an increase in semantic density (i.e. the buzzword began being used in less contexts). As the financial crisis contained three time windows, these results were compared for both pre through during and during through post.

5.4 k-NCV

The second analysis performed was a clustering of the buzzwords with the content bearing words to determine the contextual shifts associated with each event. The motivation for this was to extend the analysis to include not just how much the semantic density changed, but how the contextual usage changed. This could provide additional insight if, for example, the average cosine similarity of a buzzword remained the same pre and post event, even though the words it was being used with were different. This analysis would also allow us to examine what words are being used most similarly to our buzzwords in each time period. Using the method above, the most similar words for each buzzword in each time window were generated. To compare the topical shifts between time periods, we looked at the number of words of the top 5 k-NCV that changed pre- and post an event as well as the qualitative difference between these shifts.

6. RESULTS

In an effort to analyze the temporal shifts of presidential discourse an exploratory data analysis was conducted using the aforementioned methodologies. Within each time window we looked to find different linguistic trends, such as semantic density broadening or narrowing or the context word matches changing over time. Provided below are the results from this analysis for each analyzed time series.

6.1 9/11

As seen in Figure 1 and Figure 2, a wide array of linguistic behavior was observed for the buzzwords around the 9/11 event. As shown by the increase in semantic density, the target word "laden", meant to represent Bin Laden, experienced narrowing. The terms "wmd" and "homeland" experienced broadening shown by their decrease in semantic density. The remaining terms, a large percentage, did not show significant change in semantic density pre- and post-9/11.

Figure 1: 9/11

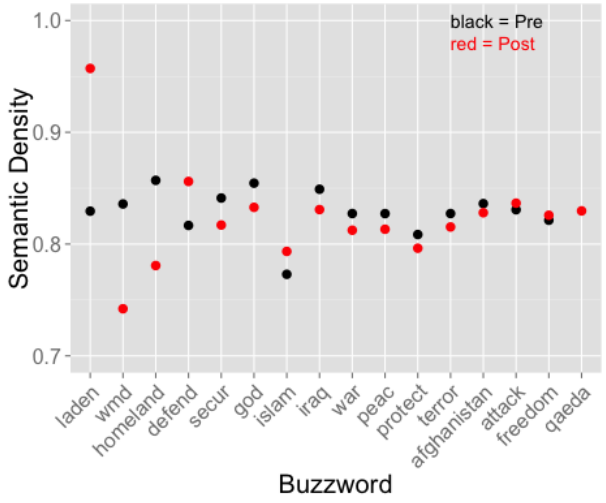
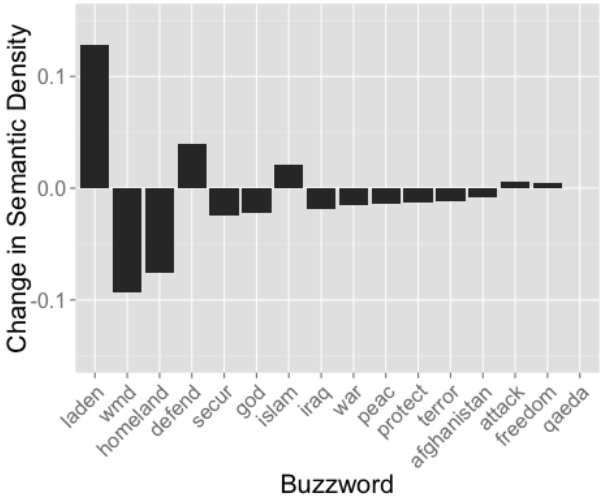
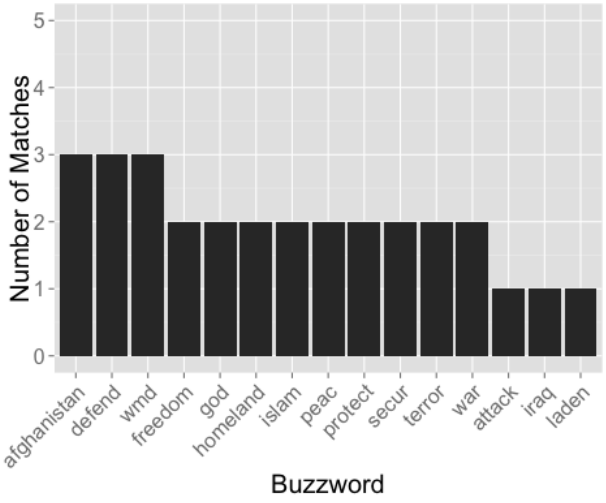


Figure 2: 9/11



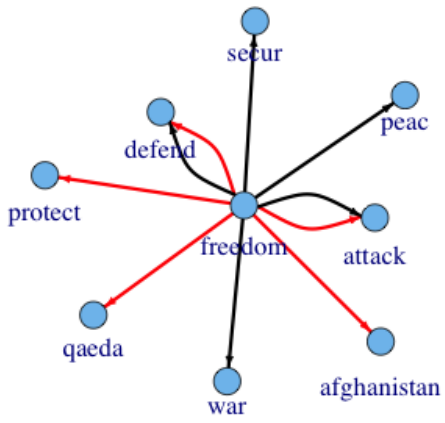
The results of the kNCV analysis, shown in Figure 3, show that majority of the target words had 2 or less matches when comparing the 5 kNCV pre and post the event. While we did not observe significant changes in the semantic density this juxtaposition indicates lack of broadening or narrowing in word usage but a shift in context.

Figure 3: 9/11



One such example of a lack of change in semantic density, but a shift in context is the target word "freedom". As shown in Figure 4, the word experienced a 3-word shift in context. Before the event, "freedom" was most closely related to "secur", "peac", and "war", but after it was most closely related to "protect", "qaeda", and "afghanistan". A logical transition given the events of 9/11.

Figure 4: 9/11



6.2 Financial Crisis

Buzzwords studied in relation to the financial crisis also showed multitude of linguistic tendencies. As shown in Figure 5 and Figure 6, minimal shifts in semantic density were seen when comparing the periods before and during the financial crisis. On the other hand, in the periods of during to after the crisis nearly all words narrowed in their context. We hypothesized that this behavior may be indicative of policy pushes by the President once blame for the crisis

and planned policy solutions, like tax credits and stimulus, took shape.

Figure 5: Financial Crisis

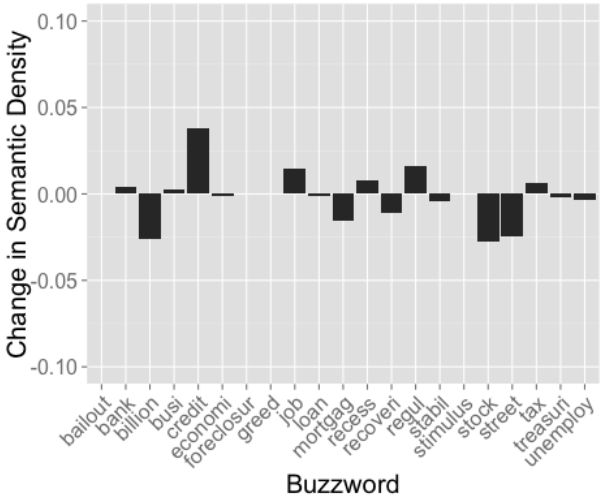
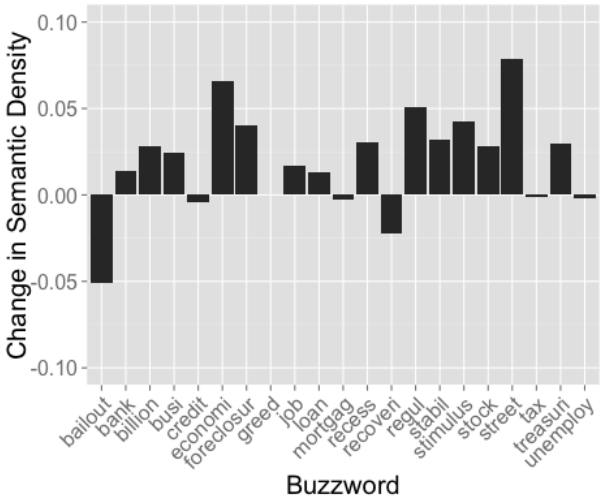


Figure 6: Financial Crisis



As seen in Figure 7 and Figure 8, the word "billion" observed massive fluctuations in its nearest topics. Prior to the crash, the word was mainly associated with topics associated with business such as "loan" and "job", however, during the crash the only remaining match was "credit" and the topics shifted to words such as "stimulus" and "economi". A similar shift occurred during and post in which the crash related words such as "credit" and "economi" shifted again back to business related words such as "bank" and "stock". Interestingly enough, the only match from during to post was "stimulus" which appeared in both kNCV.

Figure 7: Financial Crisis

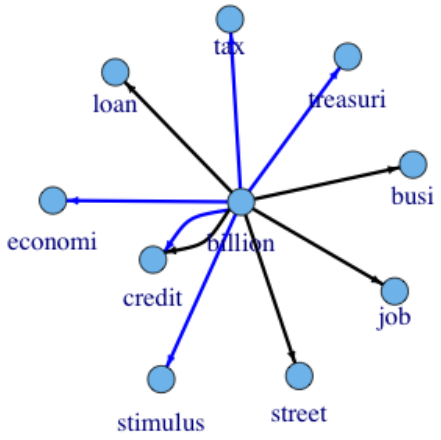
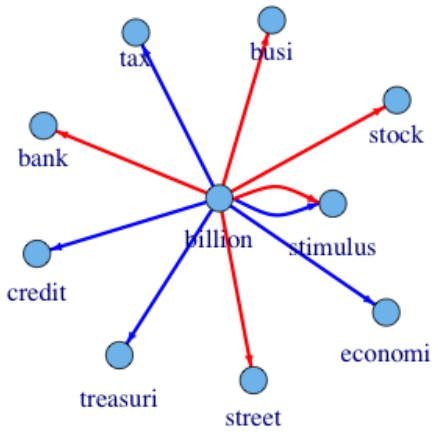


Figure 8: Financial Crisis



6.3 Sandy Hook

Notable shifts in the semantic density of buzzwords pre and post the Sandy Hook event occurred for the words "tragedi-", "background", and "mental" as shown in Figure 9 and Figure 10. All three words narrowed in their context usage, as quantified by change in semantic density. Interestingly, these three words were all most specifically related to the Sandy Hook event, thus we found this behavior a notable result. The buzzwords "god" and "kill" showed broadening.

Figure 9: Sandy Hook

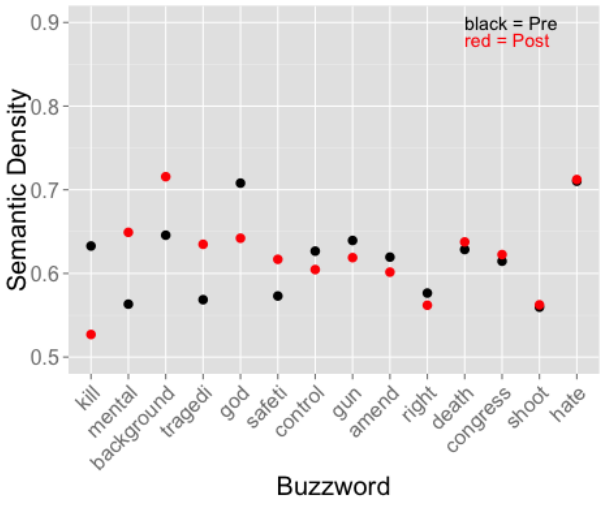
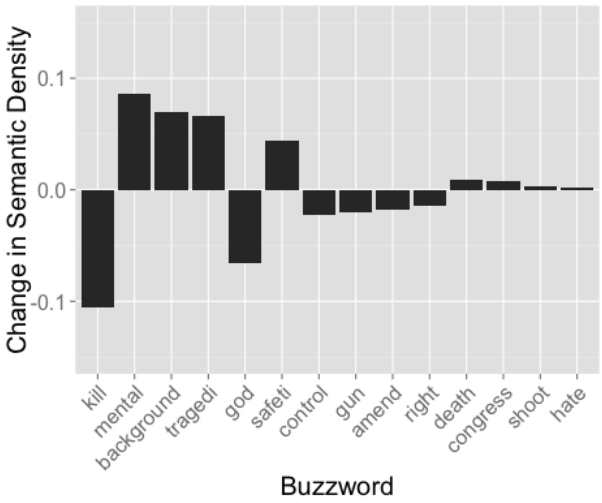
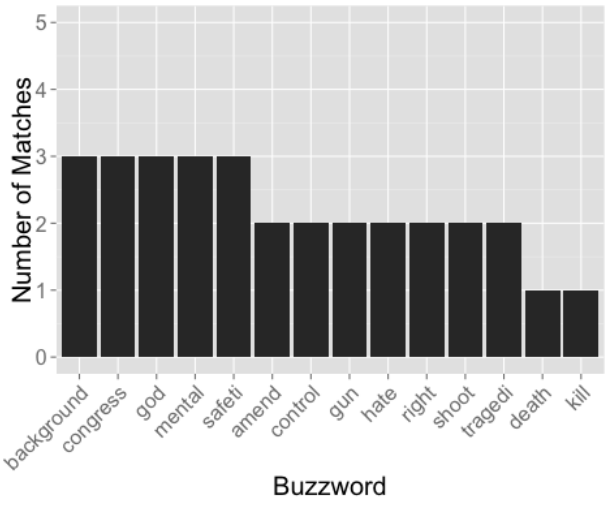


Figure 10: Sandy Hook



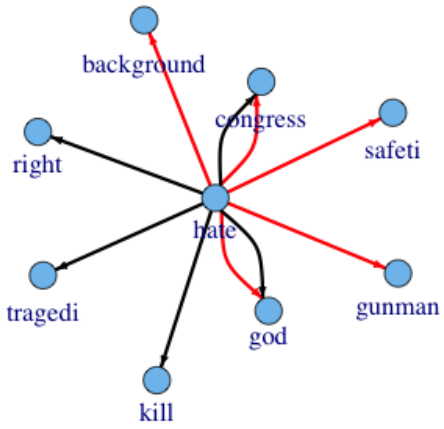
Using the results of the kNCV shown in Figure 11, we analyzed cases where the semantic density did not change, but a significant shift was seen in the kNCV. One such case was the word "hate". This word showed the smallest change in semantic density, but only 2 out of 5 kNCV matched before and after the event.

Figure 11: Sandy Hook



The network graph in Figure 12 shows the word shifts for this buzzword. The top kNCV shifted from "right", "tragedi", and "kill", general words related to hate, before the event, interestingly to "background", "safeti", and "gunman" after the event. These post-event words could be connected to policy ideas pushed after the event including gun safety and mandatory background checks for gun ownership.

Figure 12: Sandy Hook



7. CONCLUSIONS

The words used by the president are some of the most influential in the world, but despite that, they remain an underexplored resource for those looking to predict or explore the direction that the country is headed. The work in this project has shown promise in being able to tease out contextual shifts in the language of the president in response to major events. Further, this paper has applied the existing method of context vectors to a new field, the political realm.

We also developed a disruptive new approach to identifying exactly how word context changes in k-NCV, as opposed to just by how much. We feel this has lain the groundwork for interesting future research, such as analyzing more events and speakers, performing classification tasks (i.e. identifying Republican vs Democratic based on semantic density), and even forecasting changes in government policy and priority.

While these results show promise, additional work is needed. The majority of our difficulties were due to the ambiguity in natural language, such as trying to identify proper nouns (i.e. distinguishing between "street" and Wall "Street"). This was further complicated by the fact that we only used unigram tokens, which made it very difficult to identify terms such as "Bin Laden", because that was tokenized to "bin" and "laden". Additionally, while the 9/11 and Sandy Hook analyses were conducted within the same Presidency, an improvement on this methodology may seek to normalize the shifts in semantic density between speakers. This would be applicable to the Financial Crisis analysis which crossed Presidencies. In the future, we could explore ways of identifying these proper noun bigrams, either by using a bigram language model, or through a dictionary-based approach. Lastly, we acknowledge that there were only three speakers analyzed due to the nature of our dataset: Bill Clinton, George W. Bush, and Barack Obama. Future work could extend our methodology to lower-ranking members of the Executive branch, members of Congress, or political candidates.

8. REFERENCES

[1] E. Fox, M. Spies, and M. Gilat. Who was america's most well-spoken president? *Vocative*, October 2014. http://www.vocativ.com/interactive/usa/us-politics/presidential-readability/.

[2] M. Heilman. Data science on state of the union addresses. *Civis Analytics*, January 2015. https://civisanalytics.com/blog/data-science/2016/01/15/data-science-on-state-of-the-union-addresses/.

[3] O. of the Federal Register. Compilation of presidential documents. *DCPD*, pages 1–3, 2009. https://www.gpo.gov/fdsys/browse/collection.

[4] E. Sagi, S. Kaufmann, and B. Clark. Semantic density analysis: Comparing word meaning across time and phonetic space. *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 104–111, 2009. http://www.aclweb.org/anthology/W09-0214.