Using Linguistic Analyses to Understand Media Coverage of the 2016 Presidential Election

Annie V. Pham, Linguistics, Jon A. Willits, Department of Psychology
University of California, Riverside

I. INTRODUCTION

- Sentiment analysis, opinion mining, and related techniques involve the computational study of people's opinions, sentiments, attitudes, and emotions expressed in written language (Dodds et al., 2011; Kloumann et al., 2012; Liu, 2012).
- These techniques are an active research area because opinions are central to almost all human activities and are key influencers of our behavior.

Our Question

- Can we quantify how different people or topics are talked about (such as politicians like *Trump* and *Clinton*) using machine-learning methods applied to real world data, such as cable news transcripts?
- This can be answered by accessing different news sources (like CNN and Fox News) and analyzing how frequently a topic occurs, or how the positivity or negativity of the context in which it occurs.

II. METHOD

Corpus Creation

- We used LexisNexis (http://lexisnexis.com) to collect news transcripts starting from December 29, 2015 through November 8, 2016 for Fox News, CNN, and MSNBC. The corpus was divided into separate documents for each network, and for each week-long interval, resulting in 3x50 = 150 different corpus documents.
- Next, we created a program that removed unnecessary information, like document titles and page numbers.
- Then, we tokenized and preprocessed the text to remove plurality, possessives, and separated punctuation by inserting a space character, and omitted all capitalization to ensure there are no duplicates of one lexeme (e.g. *Trump* vs *trump* vs. *Trump's* should be counted as the same lexeme, see Appendix for example).

Target and Dimension Identification

- We identified a set of emotionally-loaded valence words (200 positive and 200 negative), used in a previous study that quantified positive and negative language use on social media (Willits & Seidenberg, 2017).
- These words were chosen by picking semantically-related but emotionally opposite pairs, such as *good-bad*, *win-lose*, and *truth-lie*.
- These 400 words were normatively evaluated by a set of 300 human participant raters, verifying the strength and direction of their emotional content.

Linguistic Analyses

- We created a program that uses text data taken from social media and cable news transcripts, and was able to measure several factors:
  (1) Word frequencies
  (2) Word co-occurrences
  (3) A positivity score and network bias score, based on co-occurrence counts between our target words and the positive and negative dimension words.

How Frequency and Co-occurrence were Calculated

- Frequencies were counted for our two target words (*clinton* and *trump*) and for our 400 dimension words separately for each date-corpus document.
- These counts were normalized by dividing them by the total number of words in that document, and then multiplying them by 1,000,000, resulting in a "parts per million" measure that controlled the frequencies for document size.
- Co-occurrences between target words and dimensions were counted if the two words co-occurred within a 12-word window (12 words each in both the forward and backward direction).
- Each target-dimension co-occurrence was transformed into a Pointwise Mutual Information (PMI) score, using the following formula:

$$\mathrm{pmi}(x;y) \equiv \log\frac{p(x,y)}{p(x)p(y)} = \log\frac{p(x|y)}{p(x)} = \log\frac{p(y|x)}{p(y)}.$$

- This turned the co-occurrence count into a value that was scaled by the base rate frequencies of both words, resulting in a positive number if the two words occurred more than would be expected by chance, and a negative number of the two words co-occurred less than expected by chance.

How Positivity and Bias Were Calculated

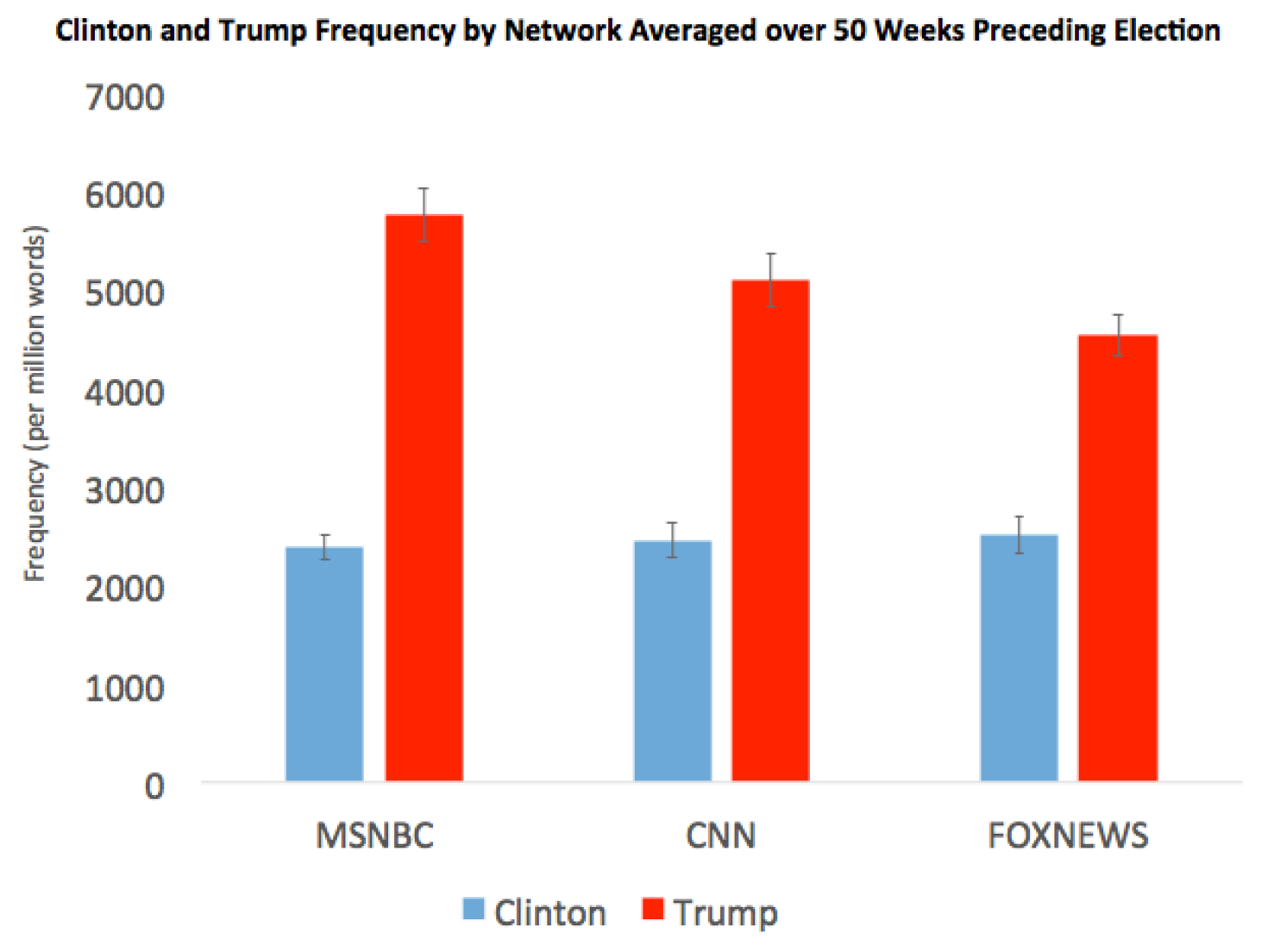
- A positivity score for each candidate, on each date, on each network was calculated by averaging their PMI scores for positive-valence and negative-valence words (with all negative words' PMI scores multiplied by -1).
- A bias score for each network on each date was calculated by subtracting Trump's positivity score from Clinton's positivity score, resulting in a positive number of the network was Clinton-biased, and a negative score if the network was Trump-biased.

III. RESULTS

Overall Frequency on Networks During the 50-week election period

- This graph to the right shows the frequencies of the candidates on the three networks, averaged across all 50 weeks.
- An 2x3 Repeated Measures ANOVA showed a significant interaction, F(2,98) = 19.62, p < 0.001. Follow-up analyses show that Trump, while significantly more frequent than Clinton on all three networks, was also more frequent on MSNBC than on CNN, and was also more frequent on CNN than on Fox News, while Clinton's frequency did not vary significantly by network.

Clinton and Trump Frequency by Network Averaged over 50 Weeks Preceding Election

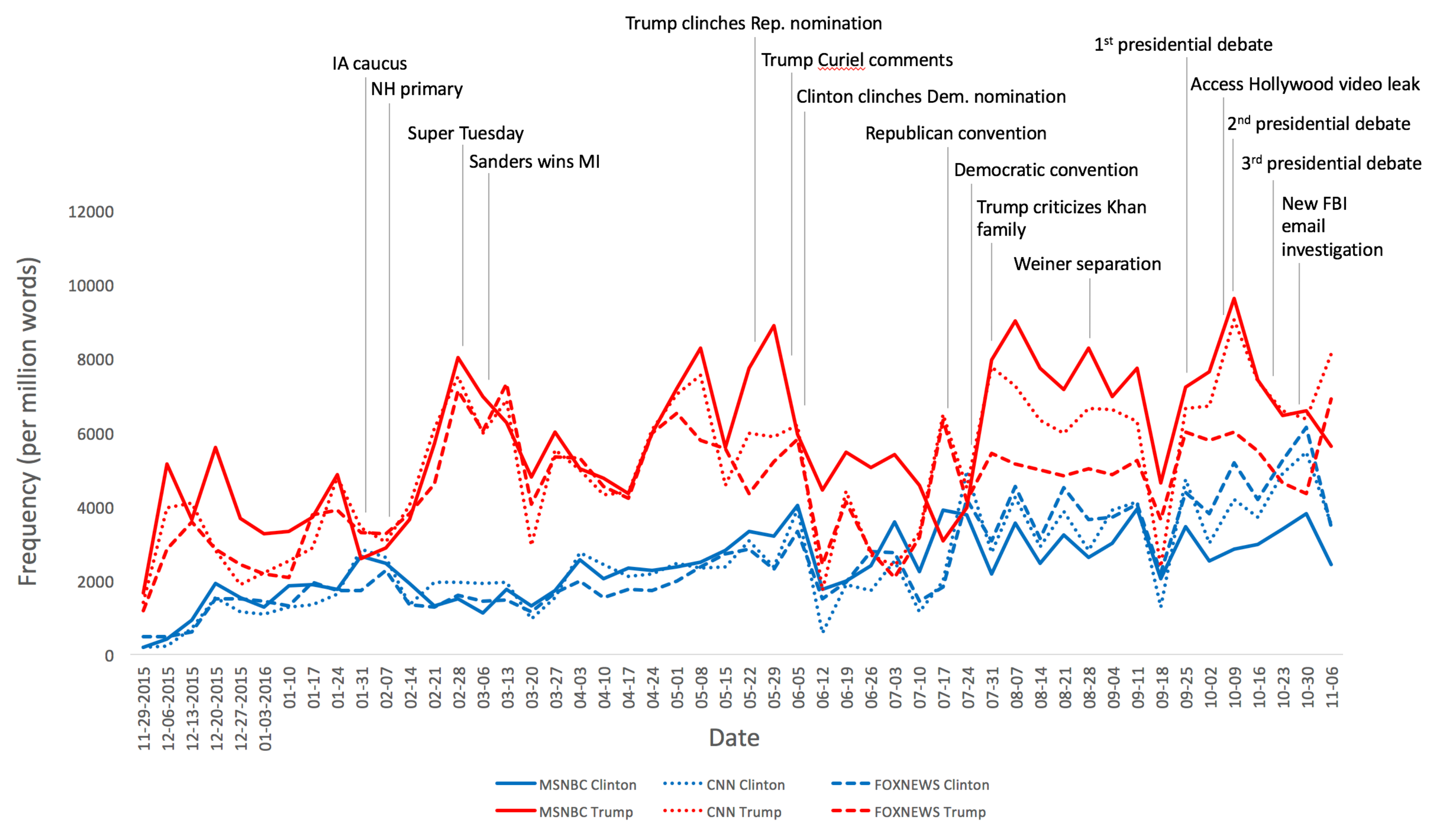


Clinton and Trump Frequency by Network Over Time

We also analyzed how frequent Clinton and Trump were on each network over time, shown in the figure below.. This graph shows us several things:

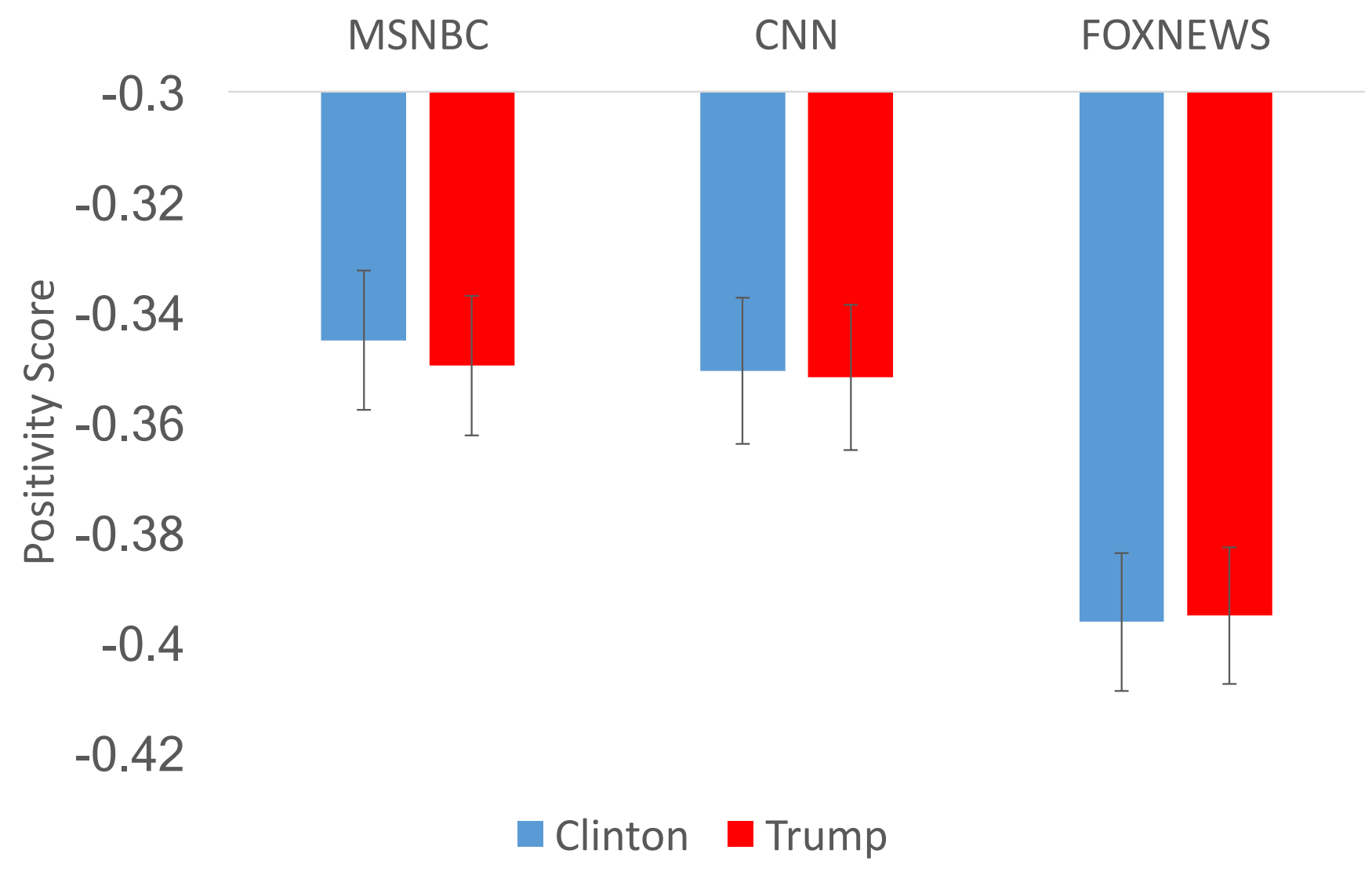
- The frequency effects noted in the bar chart above are relatively, but not perfectly, consistent over time.
- Trump's considerably pattern of higher frequency on MSNBC (and also CNN) compared to FOX News did not really emerge until after Trump had become the Republican nominee.
- Likewise, a similar, but smaller effect of Clinton being most frequent on Fox News, next most frequent on CNN, and least frequent on MSNBC, also emerged after the election.
- The frequency of the candidates was sensitive to events involving those candidates, but had different effects on different networks.

Clinton & Trump Frequency by Network



Overall Positivity on Networks During the 50-week election period

- This graph to the right shows the overall positivity of the two candidates on the three networks, averaged across all 50 weeks.
- Interestingly, scores were always negative, indicating that on average the candidates co-occurred more often than expected with negative words (relative to expected counts from base rate frequencies), and less often than expected with positive words. The size of this effect, however, varied by network.
- An 2x3 Repeated Measures ANOVA of this data showed a significant interaction, F(2,98) = 4.52, p < 0.013. Follow-up analyses show that Trump was significantly more negative than Clinton on MSNBC and CNN, while Clinton was significantly more negative than Trump on Fox News. These differences are hard to discern in the graph to the right due to the paired nature of the data, but are clear in the time series analysis below.

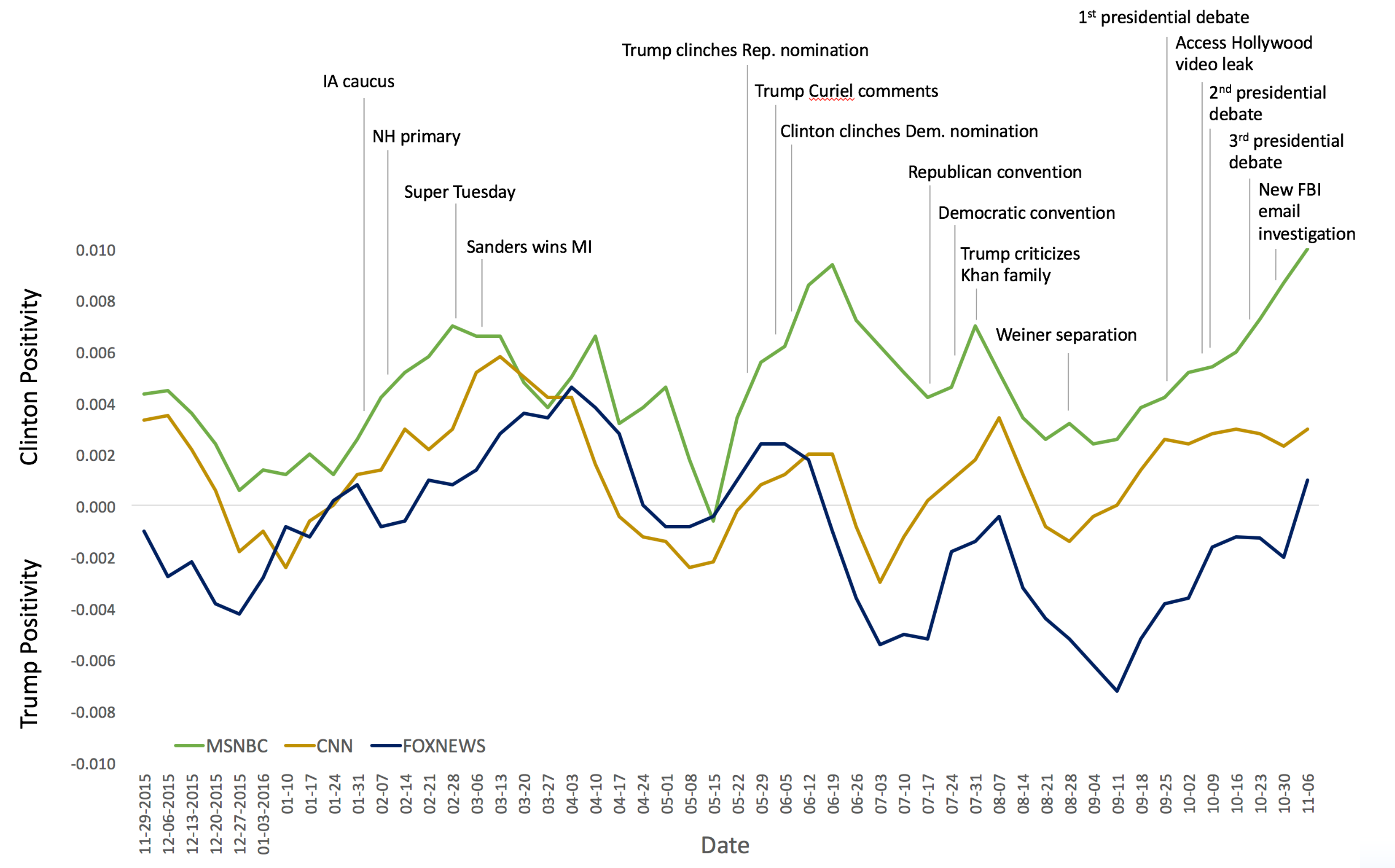


Clinton and Trump Bias Scores Over time

We also analyzed how positive Clinton and Trump were on each network over time, and converted this into a bias score (Clinton positivity – Trump positivity). This analysis shows us several things:

- As with frequency, big bias effects really emerged after the primaries were over. Even Fox News was less negative about Clinton than Trump up until Trump had clinched the nomination. At this point, a predictable partisan bias emerged.
- The positivity of the candidates was sensitive to events involving those candidates, but had different effects on different networks.

Clinton Positivity Minus Trump Positivity by Network



IV. DISCUSSION

This study showed how we can use analyses of linguistic factors like frequency and co-occurrence to describe and better understand social situations and interactions, including events like media coverage of presidential candidates. These analyses led to several important discoveries.

- Candidates' names co-occurred with emotionally-valenced words at rates that deviated significantly from what we would expect from their frequencies alone (discovered using the PMI transformation).
- By comparing the positivity or negativity of Trump and Clinton, we can measure the overall bias on each news source by subtracting the target-dimension co-occurrences.
- These co-occurrence patterns with positive and negative words lead us to a number of insights about the ways the networks covered the candidates, such as with a general tendency to be negative, predictable partisan biases, and strong differences between the primary and general election time periods.
- Important events that affected the public's opinions, such as the Trump-Curiel incident or Clinton's email investigation, had measurable correlations with changes in both frequency and positivity. However, these effects were often different for different networks.
- In future work, we will compare these analyses to more traditional sentiment analysis, to see if PMI co-occurrence counts with target words tell us the same information about how words are used, compared to when entire utterances are analyzes using sentiment analysis.
- Another future direction is to see how these analyses differ when using positive and negative words that, instead of being weighted as simply +1 and -1, instead had scores that varied by intensity (e.g. kill is more negative than slip), and also by partisan bias (e.g. peaceful is more positive for Democrats than for Republicans). These predictions of positivity can then be compared to actual metrics such as public opinion.

V. REFERENCES

- Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., & Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PloS one*, *6*(12), e26752.
- Kloumann, I. M., Danforth, C. M., Harris, K. D., Bliss, C. A., & Dodds, P. S. (2012). Positivity of the English language. *PloS one*, *7*(1), e29484.
- Liu, B. (2012). *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers.
- Willits, J. A., & Seidenberg, M. S. (2017). Political Attitudes Measured Through Language Statistics in Televised News. Manuscript under review.

VI. APPENDIX

Below is an example of the transcripts, before and after the program that tokenized and preprocessed them.

JEREMY DIAMOND, CNN POLITICS REPORTER: Absolutely, Kate. Donald Trump tonight clinching a major victory here in South Carolina. You know, this is going to help to give him some momentum going forward. I talked tonight with his South Carolina chairman Ed McMullen that told me, well, this kind of dispels all the myths. You know, Donald Trump's game ground constantly scrutinized here, as it was in the past. But Donald Trump offering some fighting words as well. Look what he had to say tonight about the pundits.

DONALD TRUMP, PRESIDENTIAL CANDIDATE: And some of the pundits, and, you know, overall fair, but not too much. But a number of the pundits said, well, if a couple of the other candidates dropped out, if you add their scores together, it's going to equal to Trump. These geniuses, they are geniuses, they don't understand that as people drop out, I'm going to get a lot of the votes also, and you don't just add them together.

absolutely , kate . donald trump tonight clinching a major victory here in south carolina . you know , this is going to help to give him some momentum going forward . i talked tonight with his south carolina chairman ed mcmullen that told me , well , this kind of dispels all the myths . you know , donald trump 's game ground constantly scrutinized here , as it was in the past . but donald trump offering some fighting words as well . look what he had to say tonight about the pundits . and some of the pundits , and , you know , overall fair , but not too much . but a number of the pundits said , well , if a couple of the other candidates dropped out , if you add their scores together , it 's going to equal to trump . these geniuses , they are geniuses , they do n't understand that as people drop out , i 'm going to get a lot of the votes also , and you do n't just add them together .

VII. CONTACT

apham024@ucr.edu, https://github.com/apham024
jon.willits@ucr.edu, http://languagelearninglab.org