ECONOMIC PERSPECTIVES

April 2016

The Candidates in Their Own Words: A Textual Analysis of 2016 Presidential Primary Debates

By Weifeng Zhong

In the 2016 election cycle, the two major parties held 20 primary debates, and the candidates spoke hundreds of thousands of words. In this paper, I turn them into "word data" and examine three characteristics of the candidates: (1) Where do the candidates stand on a spectrum of policy positions? (2) How negative are the candidates' political sentiments? (3) How effectively do the candidates' speeches deliver content? This word-data approach makes possible observations that are difficult to discover with conventional methods. For example, I find the political speeches of both Hillary Clinton and Donald Trump appear moderate in policy positions, positive in political sentiments, and effective in delivering content.

In the past decade, a big "big data" revolution has transformed how we understand the world. One of the most important sources of big data is the written and spoken words in natural language. The massive amount of textual data and the novel tools to analyze them open the door to many exciting areas of research that were previously too difficult, if not impossible.

Political science is a discipline in which big data research is particularly fruitful. Mosteller and Wallace (1963), authors of the earliest papers on textual analysis, infer who wrote the unattributed Federalist Papers by examining the style—not the content—of the writings. Treating Chinese newspapers as a dataset, Stockmann (2013) studies how media marketization influences anti-American sentiment in the Chinese public. Nielsen (2013) uses fatwas from Muslim clerics' websites to measure the level of Jihadist thought in their writings and investigate what has driven the rise of Jihadism.[1] The thrust of all this research is simple yet powerful: treat the collections of words as data

[1] See Grimmer and Stewart (2013) and Lucas et al. (2015) for more comprehensive reviews of the textual analysis method and its applications in political science.

that contain information about the texts' authors and analyze them with quantitative methods.

In this paper, I apply the techniques of textual analysis to process and analyze the 2016 US presidential primary debates. So far, the candidates have spoken hundreds of thousands of words in the 20 primary debates held by the two major parties. In this paper, I turn all of the words into data and make inferences about the candidates' characteristics that would otherwise be too difficult to discover.

Take the candidates' policy positions as an example. One might venture to say that looking at candidates' websites would suffice, but the reality often disappoints. For example, on the Donald Trump's campaign website, he lays out a comprehensive tax plan that promises big tax cuts, which President Trump would also pay for. But this tax plan is something candidate Trump seldom talks about in debates or his stump speeches—at least not as often as he talks about the wall Mexico would pay for.

Another, more rigorous method in political science is to estimate policy positions using legislators' roll-call votes—an industry standard invented by then–Carnegie Mellon professors Keith T. Poole and Howard Rosenthal (1985). For candidates concurrently serving in Congress, this method tells us very well where they stand.[2] Ted Cruz, for example, is very conservative; Marco Rubio less so; and Bernie Sanders very liberal. However, some candidates served in Congress only in the relatively distant past, such as Hillary Clinton (senator from New York, 2001–09) and John Kasich (representative from Ohio, 1983–2001). Some other candidates have never served in Congress, such as Jeb Bush, Ben Carson, and—most importantly—Donald Trump. For these candidates, conventional methods tell us little.

Fortunately, all candidates talk, and by now they have talked repeatedly about various policy issues in presidential primary debates. Moreover, candidates often try to differentiate themselves from their rivals on the debate stage by making statements in a common context using similar lexicons. This makes publicly available transcripts of debates a valuable resource for systematically comparing candidates with each other and over time.

In this paper, I examine three characteristics of the presidential candidates by analyzing their debate transcripts: (1) where the candidates stand on a spectrum of policy positions, (2) how negative their political sentiments are, and (3) how effectively their speeches deliver content. Among other findings, I show that the political speeches of Hillary Clinton and—perhaps more surprisingly—Donald Trump have the following in common: their policy positions appear quite moderate, their political sentiments are not too negative, and their simple language delivers content rather effectively.

Data and Methodology

The raw data used in this paper are the full debate transcripts in the 2016 election cycle available at the University of California, Santa Barbara's American Presidency Project (see Peters and Woolley, 2016). Included are the eight debates in the Democratic field held through March 9 and the 12 "main-stage" debates on the Republican side held through March 10. In each debate, each candidate spoke thousands of words. For this paper, I consolidate everything a candidate has said in a debate into a single, long statement, regardless of whether the sentences are answers to moderators' questions or responses to other candidates' comments.

Natural language is complex. Therefore, standard processing steps identified in the textual analysis literature are used to reduce the complexity of candidates' statements. The most substantial step is to discard the order in which words appear and to think of each statement as a list of counts of how many times a candidate has used certain words. For example, in the most recent Republican debate, Donald Trump's statement contains 9 uses of for the word *China*, 26 of *country*, 14 of *great*, and so forth. There is no question that word order matters to the substantive meaning of sentences and that looking at only word counts therefore loses some information. For example, Donald Trump has

[2] See Poole and Rosenthal's website, http://voteview.com/, for the most up-to-date analysis.

never said “China is a great country” in any debate, but that fact is not reflected in his word counts. Nevertheless, for such tasks as estimating policy positions and analyzing sentiments, this drastic reduction of natural language performs surprisingly well (e.g., Hopkins and King, 2010).

The candidates’ vocabularies are further simplified by removing common words that play mostly functional roles and convey little meaning (e.g., articles, prepositions, and pronouns) and stemming the remaining, more substantive words. Stemming refers to the process of stripping the suffixes off words and keeping only their roots.[3] The word stem *presid*, for instance, appears in Hillary Clinton’s statement 29 times in the most recent Democratic debate, corresponding to the original words *president*, *presidents*, and *presidency* that convey similar meanings.

Throughout, I refer to a processed statement by a candidate as that candidate’s “text” in the debate, and the analyses in this paper are based on these texts.

Policy Positions: How Conservative Is Donald Trump, According to Donald Trump?

The first question is: where do candidates stand on a political spectrum from the most liberal position to the most conservative one? Donald Trump, for example, often proclaims himself a “commonsense” conservative. But how far right is he relative to Bernie Sanders, Marco Rubio, and Ted Cruz, whose positions are well known from their roll-call votes?

I answer this question using Wordscores, a supervised scaling method developed in the groundbreaking work by Laver et al. (2003). The idea of Wordscores is straightforward: the analyst feeds the algorithm with reference texts from candidates at known policy positions, and then the algorithm “learns” how to situate all other candidates by comparing their texts with the reference texts. I provide a nontechnical summary of the method in this section. More technical details can be found in Laver et al. (2003).

Method. Three reference candidates are selected for this study: Bernie Sanders (assumed with policy position –1), Marco Rubio (position 0.6), and Ted Cruz (position 1).[4] For each candidate, I concatenate his texts across *all* debates and treat the compilation as a reference text. The main reason for the concatenation across debates is to include as many words as possible used by each reference candidate so that the algorithm has a rich set of lexicons to compare with other texts. An added benefit is that the reference candidates themselves can be scored as well. For example, even compared with the overall image of his concatenated text, Ted Cruz’s debate performances may exhibit different degrees of conservatism at different times.

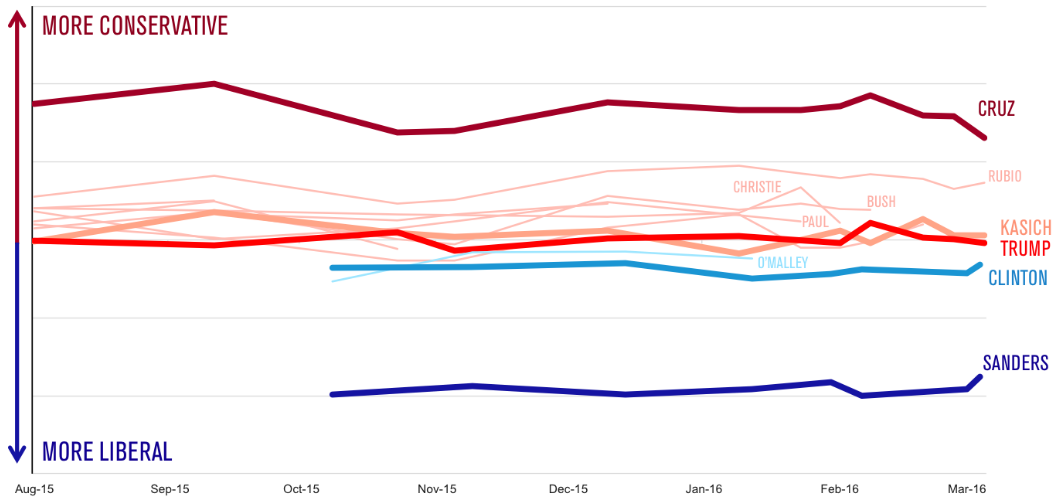
The key part of the algorithm is to generate “conservatism scores” for the word stems in the reference texts. If a word stem is used by two GOP candidates—especially by the very conservative Ted Cruz—at a relatively higher rate than by the very liberal Bernie Sanders, then we classify the word stem as more conservative and assign a higher score to it. The score is lower if the opposite is true. The interpretation of a higher conservatism score is predictive: if we see a conservative word stem appear more frequently in a new unscored text, then it is more likely that the new text comes from a conservative candidate. For this very reason, the algorithm works best if the reference candidates are known to be at extreme positions from the start because the words they use have more predictive power.

Because many different words are used in a single text, I measure the overall policy position of an unscored text by the weighted average of the scores of all word stems it contains. Theoretically, a new text could have a position even more extreme than Bernie Sanders’ or Ted Cruz’s text. For example, if a candidate repeated a very liberal (or a very conservative) word 10,000 times on the debate stage, it would yield a very low (or very high) average score. In other words, the positions of

[3] The stemming method used in this paper is called the Snowball algorithm (http://snowballstem.org/), a widely used method invented by computer scientist Martin F. Porter (1980).

[4] The parameters are chosen to be consistent with Poole and Rosenthal’s roll-call voting analysis up to the 113th Congress. See http://voteview.com/ for details.

Figure 1. 2016 Candidates' Policy Positions



Source: Author's analysis of the transcripts of presidential primary debates.

unscored texts are not bounded by the reference texts a priori.

Results. Figure 1 shows the main results of this analysis, based on the candidates' own words. Highlighted in different colors is how the policy positions of the five running candidates in the two parties have changed over the course of multiple debates. Candidates who dropped out before March are not highlighted.

First, John Kasich *and* Donald Trump appear among the *least* conservative candidates in the Republican field from the beginning. Falling in the moderate territory, the two candidates' positions are not far from Hillary Clinton's. This is not a surprise on the part of John Kasich—or Hillary Clinton for that matter—who has been widely characterized as a Republican liberals can like.[5] Donald Trump, however, turns out to be just as moderate as John Kasich.[6] Moreover, for the most part, the positions of these two GOP candidates have been tightly intertwined. To the extent that they are competing for the same group of moderate voters, Donald Trump has apparently won the contest lopsidedly.[7]

Second, the policy positions of Republican candidates tend to vary more over time than those of the Democratic candidates. In the CNBC debate on October 28, for example, both Marco Rubio and Ted Cruz were more than 30 percent *less* conservative than they were in the previous debate.[8] In contrast, Bernie Sander was usually within 10 percent of his most liberal performance throughout. The larger fluctuation in positions on

[5] For example, see the endorsement of John Kasich by *New York Times*' Editorial Board (2016).

[6] This analysis does not rule out the possibility that Donald Trump's position is only moderate on average. There are different ways to appear moderate in a debate. Being consistently moderate in every issue is one way. Making both conservative and liberal remarks in the same debate is another. For example, Donald Trump is known for his right-wing position on illegal immigration as well as having said he would not change entitlements, which all other GOP candidates have pledged to reform. Hence, his inconsistency on different issues may have driven his moderate score in Figure 1.

[7] The result would be similar if the analysis was done with only two reference candidates: Bernie Sanders and Ted Cruz. In that case, Marco Rubio would still be the second most conservative overall, and John Kasich and Donald Trump would still be intertwining in the moderate territory. However, Marco Rubio is included as a reference candidate because his positions are relatively well-known based on roll-call votes.

[8] One may be tempted to attribute the "unconservative" performance of Republican candidates on October 28 to the widely criticized CNBC moderators. However, in the next debate, held by the Fox Business Network and *Wall Street Journal* on November 10, the Republican candidates were still just about as unconservative.

the Republican side can be explained by the crowded field. The more rivals a candidate faces, the more likely that candidate will try to position differently to stand out on national television. The GOP field as a whole, then, swings more with all the candidates trying to do the same.

Third, as the field narrows over time, extreme candidates have moved toward the center of the spectrum. In the most recent debate, for instance, Ted Cruz was only 65 percent as conservative as he was at his peak, whereas Bernie Sanders was at 88 percent of his most liberal moment. On the Republican side, it would be especially interesting to see if Ted Cruz moves even closer to the moderate territory if there are any more GOP debates after Marco Rubio's exit. Ted Cruz's recent effort to win over Rubio supporters suggests that the Texas senator may become even more moderate (see, e.g., Glueck, 2016).

Finally, to validate the method of this analysis, I show in Table 1 that Wordscores picks up important policy issues that define and divide the

Table 1. Most Frequently Used Liberal and Conservative Words

Liberal			Conservative		
Rank	**Word Stem**	**Note**	**Rank**	**Word Stem**	**Note**
1	wealth		1	amnesti	
2	almost		2	everyon	
3	financ		3	obamacar	
4	vermont		4	flat	flat tax
5	flint	Flint, Michigan	5	marco	
6	revolut		6	chief	
7	unemploy		7	note	
8	infrastructur		8	command	
9	fossil		9	liberti	
10	virtual		10	liber	
11	overthrow		11	ir	IRS
12	anderson	Anderson Copper	12	religi	
13	dictat		13	texa	Texas
14	michigan		14	launch	
15	latino		15	abolish	
16	goldman	Goldman Sachs	16	author	authority
17	sach		17	parenthood	
18	crumbl		18	reid	Harry Reid
19	shut	companies shutting down	19	target	
20	aggress	act aggressively	20	politician	

Note: Word stems with scores no higher than –0.85 are called liberal word stems. The cutoff for conservative word stems is 0.85. The word stems in the table are sorted in descending order by how often they are used by the reference candidates.
Source: Author's analysis of the transcripts of presidential primary debates.

field of the presidential contest. Listed in the table are the 20 most frequently used liberal and conservative word stems. To appear more liberal, Democratic candidates focus on wealth and income inequality, large financial institutions, unemployment, Latinos, and so forth. In contrast, GOP candidates emphasize immigration, Obamacare, taxes, abortion, and so forth to appear conservative. This stark contrast in word choice is largely consistent with the polarization in American politics throughout the past several decades (e.g., McCarty et al., 2006).

Political Sentiments: Does a Candidate Need to Appear Angry to Get Angry Votes?

The political sentiment of the American public is often characterized as increasingly negative, with more people feeling that they are becoming worse off and that the country is going in the wrong direction. Given an angry and frustrated electorate, do candidates need to appear angry and frustrated as well to get their votes? In this section, I conduct a sentiment analysis of candidates' debate performance to answer this question.

Method. I use a dictionary method to study the sentiment in debates. The method, which is most intuitive in automated textual analysis, measures the tone of a text's author by the rate at which certain positive and negative words—defined by the dictionary—appear in the text.[9] The dictionary consists of a list of positive words and another list of negative words, without measuring the intensity of each word's sentiment. Therefore, an unscored new word could be positive, negative, or neutral if it is not in either list. For each debate transcript, we classify the candidate's overall political sentiment as more negative if he or she uses more negative words and fewer positive words.[10]

Many dictionaries exist for conducting sentiment analysis. Here I use the sentiment dictionary developed by University of Pittsburgh's Multi-Perspective Question Answering project, which is freely available and is one of the most widely used in the sentiment analysis literature.[11]

Results. The main results are summarized in Figure 2, which is organized to be comparable to Figure 1. Democratic candidates are graphed in the lower panel, and Republican candidates in the upper one.

First, Democratic candidates appear less negative than GOP candidates. In particular, Hillary Clinton is the most positive of the candidates in both parties. This finding can be easily explained by the fact that the Democratic Party is the incumbent party and that Hillary Clinton is the candidate most closely associated with the Obama administration. The difference in tone between the two parties was perhaps most pronounced when the candidates on both sides talked about President Obama and the Affordable Care Act on the debate stage, with Ted Cruz fiercely criticizing it and Hillary Clinton passionately praising it.

Second, in both parties, the candidates at more extreme policy positions, as identified earlier, tend to be more negative than candidates at less extreme positions. Bernie Sanders, for example, is more negative than Hillary Clinton, and Ted Cruz is more negative than John Kasich. This pattern is clearly driven by the need of extreme candidates to harshly criticize the status quo the current government represents, whether it be the "establishment politics and establishment economics" coined by Bernie Sanders or the "Washington cartel" seen by Ted Cruz.

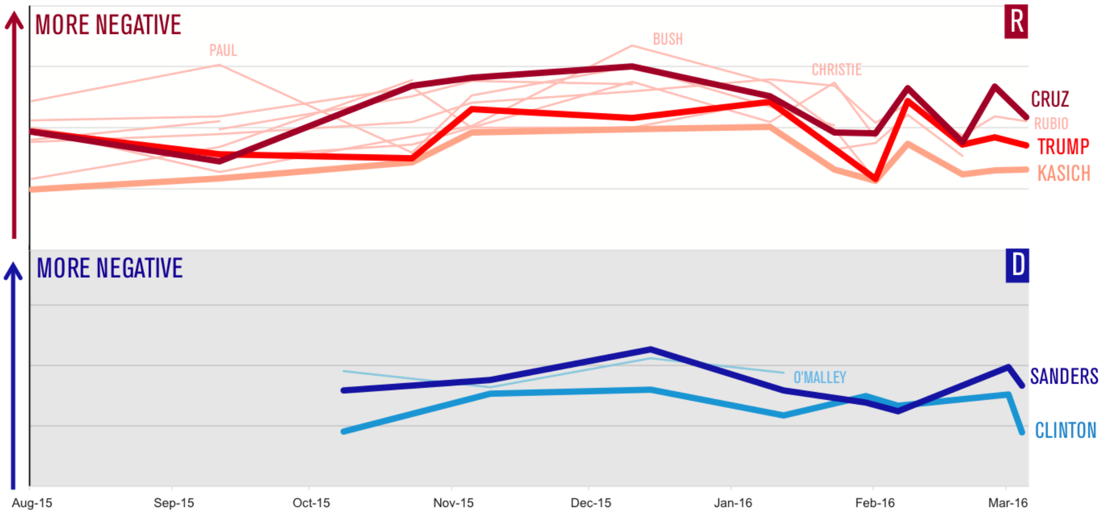
Most notably, Donald Trump—a very "unconservative" candidate—is also rather positive in the Republican field. Widely believed to be targeting American voters who are angry and frustrated about the status quo, Donald Trump has been winning their votes with a positive tone. In the following excerpt from the CNBC debate on October 28, Donald Trump responded to John

[9] Wordscores, used in Section 2, is itself a dictionary method. The dictionary is developed from the word stems in the reference texts, each with its own conservatism score. All the new texts are then scored by this dictionary.

[10] When counting positive and negative words, I also take into account whether they are used in a negation. For example, if a candidate says "the country is not great anymore,"—great, a positive word, is negated by not—it is counted as a negative expression, instead of a positive one.

[11] The dictionary's positive and negative words are available as separate lists at http://www.unc.edu/~ncaren/haphazard/positive.txt and http://www.unc.edu/~ncaren/haphazard/negative.txt, prepared by Neal Caren at the University of North Carolina, Chapel Hill.

Figure 2. 2016 Candidates' Political Sentiments



Source: Author's analysis of the transcripts of presidential primary debates.

Kasich's criticism that dealing with waste, fraud, and abuse alone will not solve the fiscal problem. In this response, although he used many positive words, *nasty* is the only negative one picked up by the dictionary:

> John got lucky with a thing called fracking. . . . Believe me, that is why Ohio is doing well. . . .
>
> [He] was a managing general partner at Lehman Brothers when it went down the tubes. . . .
>
> [H]e was so nice. He was such a nice guy. And he said, oh, I'm never going to attack. But then his poll numbers tanked. He has got—that is why he is on the end. And he got *nasty*. (Peters and Woolley, 2016)

Donald Trump seems to be aware of an important point: he does not need to appear angry on national television merely because he is courting angry voters.

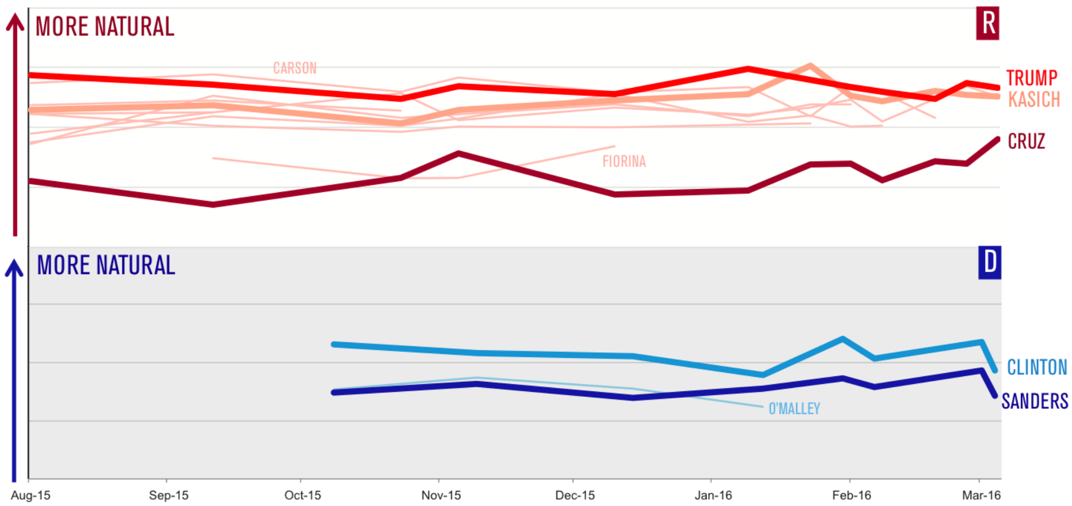
Speech Effectiveness: How Well Do Candidates Deliver Content?

A lot of attention has been paid to how much time a candidate has spoken in each debate. However, not every minute a candidate speaks is effective in delivering content to the audience. In this section, I examine the speech effectiveness in each candidate's transcripts using a crude yet informative measure.

Method. As noted earlier, when processing the raw transcripts of debates, certain common words that are mostly functional and containing little meaning—such as articles, prepositions, and pronouns—are removed prior to the analysis of textual substance. These words are called "stop words" in the textual analysis literature.[12] Although not substantive themselves, stop words exist in natural language for a good reason: they attest to the use of substantive words and facilitate the delivery of actual content. A text with very few stop words, in contrast, would be too complex for the

[12] There are many lists of stop words. Throughout this paper, I use the stop-word list in Python's Natural Language Toolkit (NLTK). For more details about it, see Chapter 2 of Bird et al. (2016). For an example of more aggressive stop-word lists (i.e., lists that may contain words with substantive meaning), see the list created by Lewis et al. (2004), which can be accessed at http://jmlr.org/papers/volume5/lewis04a/a11-smart-stop-list/english.stop. Nevertheless, the main results of this paper hold up with different lists.

Figure 3. 2016 Candidates' Speech Patterns



Source: Author's analysis of the transcripts of presidential primary debates.

audience to understand. Therefore, to measure the effectiveness of candidates' speeches, a crude approximation is to compute the fraction of stop words that appear in a transcript. If that fraction is larger for a candidate's transcript, we classify this candidate's speech as more effective in delivering content.

In textual analysis, the stop-word ratio has been used to capture the readability or quality of documents (e.g., Kanungo and Orr, 2009; Bendersky et al., 2011). Bendersky et al., for example, show that Wikipedia entries—which are supposedly very readable and informative— have a higher stop-word ratio than ordinary webpages.

Results. Figure 3 summarizes the comparison in speech effectiveness among candidates across debates.

On the Democratic side, Hillary Clinton's speech is more effective than that of Bernie Sanders. Perhaps more surprisingly, Donald Trump is more effective than his GOP rivals in delivering his content to the audience. This may have been a strategy Donald Trump has adopted: to talk to a "low-information" audience, a wise politician would "dial down" the language to deliver content.

Ordering candidates by their speech effectiveness would be very similar if the effectiveness was measured by how *few* characters each of their words has—shorter words are easier to understand. The result (not graphed in this paper) shows that Donald Trump has 3.4 characters per word on average, while Ted Cruz has 3.9. The number of characters per word is 3.8 for Hillary Clinton and lower than 3.9 for Bernie Sanders. The ordering of candidates is also consistent with existing analyses of the grade level required for a person to understand the candidates' speeches (e.g., Viser, 2015). Donald Trump, for example, speaks at a very comprehensible fourth-grade level, compared with Ted Cruz's ninth-grade level.

Another way to interpret the result is to compare the speech effectiveness by candidates' political background (or the lack thereof). Politicians with legislative experience are the least effective, with an average score of 0.54. Politicians with executive experience[13] are more effective, with an average score of 0.56. Nonpoliticians' speeches are the most effective, with an average score of 0.58. An obvious explanation of this difference is that legislative language is the most complex and rigorous and, hence, the hardest to understand. In contrast, government executives, businessmen and

13 If a candidate has served in both the legislative and executive branches, the candidate's most recent experience is used for the categorization. For example, John Kasich, who was a US Representative but is currently a governor, falls into the second group.

businesswomen, and the like interact with ordinary people more often and therefore tend to use the simpler, daily language.

At least so far, this strategy of using simple language seems to have served Hillary Clinton and Donald Trump quite well in the race.

Conclusion

In the current election cycle, the candidates have spoken hundreds of thousands of words in the 20 primary debates held by the two major parties. This paper turns their statements into word data and examines candidates' characteristics that are difficult to discern with conventional methods.

Among other things, this paper shows that Hillary Clinton and Donald Trump—the frontrunners in the two parties—use similar political speech in comparison with their party rivals: (1) their policy positions are quite moderate, (2) their political sentiments are not too negative, and (3) their simple language delivers content to the audience rather effectively. If they become their parties' nominees, we should expect these features to become even stronger in the general-election debates.

About the Authors

Weifeng Zhong (weifeng.zhong@aei.org) is a research fellow in economic policy studies at the American Enterprise Institute.

References

Bendersky, Michael, W. Bruce Croft, and Yanlei Diao. 2011. "Quality-Biased Ranking of Web Documents." *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. February 9–12: 95–104.

Bird, Steven, Ewan Klein, and Edward Loper. 2016. *Natural Language Processing with Python*. Manuscript. http://www.nltk.org/book/.

Editorial Board. 2016. "A Chance to Reset the Republican Race." *New York Times*, January 30. http://www.nytimes.com/2016/01/31/opinion/sunday/a-chance-to-reset-the-republican-race.html.

Glueck, Katie. 2016. "Inside Cruz's Bid to Conquer Rubioworld." *Politico*. March 19. http://www.politico.com/story/2016/03/ted-cruz-marco-rubio-donors-supporters-220969.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–97.

Hopkins, Daniel J., and Gary King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54 (1): 229–47.

Kanungo, Tapas, and David Orr. 2009. "Predicting the Readability of Short Web Summaries." *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. February 9–12: 202–11.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *The American Political Science Review* 97 (2): 311–31.

Lewis, David D., Yiming Yang, Tony G. Rose, and Fan Li. 2004. "RCV1: A New Benchmark Collection for Text Categorization Research." *Journal of Machine Learning Research* 5: 361–97.

Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley. 2015. "Computer-Assisted Text Analysis for Comparative Politics." *Political Analysis* 23 (2): 254–77.

McCarty, Nolan, Keith T. Poole, and Howard Rosenthal. 2006. *Polarized America: The Dance of Ideology and Unequal Riches*. Cambridge, MA: MIT Press.

Mosteller, Frederick, and David L. Wallace. 1963. "Inference in an Authorship Problem." *Journal of the American Statistical Association* 58 (302): 275–309.

Nielsen, Richard Alexander. 2013. *The Lonely Jihadist: Weak Networks and the Radicalization of Muslim Clerics*. Doctoral dissertation, Harvard University.

Peters, Gerhard, and John T. Woolley. 2016. "Presidential Debates." *American Presidency Project*. http://www.presidency.ucsb.edu/debates.php.

Poole, Keith T., and Howard Rosenthal. 1985. "A Spatial Model for Legislative Roll Call Analysis." *American Journal of Political Science* 29 (2): 357–84.

Porter, Martin F. 1980. "An Algorithm for Suffix Stripping." *Program* 3 (14): 130–7.

Stockmann, Daniela. 2013. *Media Commercialization and Authoritarian Rule in China*. New York, NY: Cambridge University Press.

Viser, Matt. 2015. "For Presidential Hopefuls, Simpler Language Resonates." *Boston Globe*. October 20. https://www.bostonglobe.com/news/politics/2015/10/20/donald-trump-and-ben-carson-speak-grade-school-level-that-today-voters-can-quickly-grasp/LUCBY6uwQAxiLvvXbVTSUN/story.html.

AEI Economic Perspectives is a publication of the American Enterprise Institute for Public Policy Research (AEI), a nonpartisan, not-for-profit, 501(c)(3) educational organization. The views expressed in AEI publications are those of the authors. AEI does not take institutional positions on any issues.