Big Social Data Analytics in Journalism and Mass Communication:

Comparing Dictionary-based Text Analysis and Unsupervised Topic Modeling

A pre-print of the article that appeared in *Journalism & Mass Communication Quarterly, 93*(2), 332-359. https://doi.org/10.1177/1077699016639231

Guo, L., **Vargo, C.,** Pan, Z., & Ding, W. (2016)

Abstract

This paper presents an empirical study that investigated and compared two "big data" text analysis methods: dictionary-based analysis, perhaps the most popular automated analysis approach in social science research; and unsupervised topic modeling (i.e., LDA analysis), one of the most widely used algorithms in the field of computer science and engineering. By applying two "big data" methods to make sense of the same dataset—77 million tweets about the 2012 U.S. presidential election, the study provides a starting point for scholars to evaluate the efficacy and validity of different computer-assisted methods for conducting journalism and mass communication research, especially in the area of political communication.

Keywords

computer-assisted content analysis, unsupervised machine learning, topic modeling, political communication, Twitter

McQuail notes that "the entire study of mass communication is based on the premise that the media have significant effects" (1994, p. 327). However, whether the "premise" still holds true in this transforming media environment remains a question. The latest Gallup polls show that Americans' confidence in major news media platforms continued to decline in the past few years (Morales, 2012). Instead, people now turn to a wide variety of alternative media outlets such as blogs and social networking sites for news and information, and on occasion become "citizen journalists" or "produsers" themselves (e.g., Napoli, 2011; J. Rosen, 2008; S. Rosen, 2010). As a response to these changes, it has become a major priority for journalism and mass communication researchers to answer questions like: Do traditional news media still have significant effects on public opinion? What other media platforms determine what the public thinks? Can public conversations on emerging media platforms potentially set the media agenda?

The investigation of all the above questions requires the empirical analysis of *content* in different media outlets and of public opinion. In the pre-Internet era when media content was limited to newspaper articles and broadcast news transcripts, a manual content analysis was sufficient to detect topics, attributes or frames inherent in the media text. As for public opinion, research methods such as interviews and surveys are considered ideal to extract beliefs and attitudes from the public's mind. Today, the widespread availability and accessibility of a large volume of media and public opinion data on Twitter, Facebook, YouTube, Reddit, and other new communication channels open the door for unprecedented research opportunities. However with these new possibilities come new challenges. The size of the data—e.g., millions or even billions of units of analysis—is beyond what traditional social science research methods can handle.

It is in this context that the investigation of "big data" analytics and its direct application in journalism and mass communication research is particularly crucial and timely.

This paper presents a methodological exploration of computational methods for processing big *social* data—text-based big data collected from various social networking sites. While no hard and fast rule defines what size is awarded the status of "big data," here we assume millions of units. Researchers in the field of computer science and engineering have developed a number of algorithms to automate the processing of large-scale text analysis during the past decade. However, the question remains whether these algorithms can generate valid and reliable results, or the degree to which those results "make sense" and are of sufficient rigor for journalism and communication contexts. Our knowledge is also limited as to which method(s), among a wide range of choices, can produce the most meaningful output while remaining cost-effective.

As an attempt to explore big social data analytics for the purpose of journalism and mass communication research, this study empirically examines two automated text analysis approaches: dictionary-based text analysis and unsupervised topic modeling. Each approach is used respectively to discover salient topics in 77 million tweets collected during the 2012 U.S. presidential election. The results generated by each method are compared with each other and with a sample of human evaluations. The strength and weakness of each approach is discussed. Overall, this paper hopes to provide a platform for scholars to assess which computational method is more beneficial for certain research scenarios.

Content Analysis in Communication Research

Manual Content Analysis

Manual content analysis is one of the most popular quantitative research methods in the field of journalism and mass communication. Some 30 percent of all journalism and mass communication research relied on manual content analysis as its main method of investigation (Kamhawi & Weaver, 2003). Despite the large number of content analysis studies, the target of analysis has traditionally been news media content. As of 1997, newspaper (46.7%) was the predominant medium for content analysis while another quarter (24.3%) focused on television transcripts (Riffe & Freitag, 1997). The advent of the Internet has allowed a vast expansion of the types of media that mass communication researchers have content-analyzed. Websites, blogs, Twitter, Facebook, and other social platforms are now emerging as large repositories of textual information rife for the picking (Lacy, Duffy, Riffe, Thorson, & Fleming, 2010; Leccese, 2009; Xiang, 2013).

At its core, manual content analysis is the process of categorizing data based on human input to answer some greater research question about the data (Riffe, Lacy, & Fico, 2014). As examples, a manual content analysis approach can answer research questions like: What is the most salient topic in the news coverage of a political election? How often are government officials mentioned in the reporting of social protests? In practice, researchers start by designing a codebook with predefined categories (e.g., a list of issues, a list of personal qualifications). To decide the topic or attribute categories for analysis, researchers usually use deduction—e.g., review previous literature—or/and induction— e.g., review a representative sample of text—to discover which topics or attributes are the most salient. Human coders then classify the texts in terms of these categories.

In order to limit the subjectivity of individual human coders, careful training of coders and several rounds of intercoder reliability tests are performed prior to and sometimes after the analysis (Krippendorff, 2004). Perfect intercoder agreement is, however, impossible and thus a certain degree of discrepancy between coders is often tolerated.

Manual content analysis is beneficial in that human coders can easily detect the nuances and complexities within the text (e.g., sarcasm) that are very hard for computers to detect without advanced methodology (e.g. natural language processing). However, this traditional method is expensive and time consuming. In addition, human errors are inevitable.

As for the data size, traditional content analysis has dealt with datasets that may be considered "big" in nature through the use of a systematic sampling procedure. In the literature, a debate exists as to exactly "how much is enough" (Connolly-Ahern, Ahern, & Bortree, 2009; Hester & Dougall, 2007; Luke, Caburnay, & Cohen, 2011). The authors of these papers suggest that sample size varies by subject domain, by media being sampled and by variable type being analyzed. While Riffe, Fico and Lacy summarize the literature and offer straightforward sampling suggestions for traditional media content, they concede that there is a "difficulty of creating a sampling frame" for big data (2014, p. 93). As a result, no clear sampling guidelines exist for big data sets that could potentially involve as many as one billion units spanning days to years (Goel, Anderson, Hofman, & Watts, 2013). At current, it is almost impossible to rely on human coding alone to interpret big social data in a systematic manner.

Computer-assisted Text Analysis

Outside of sampling populations, mass communication researchers have turned to computers to automate content analysis tasks (West, 2001). Lewis, Zamith and Hermida (2013) were early to recognize the value of computational methods in processing big social data, saying that these methods "in theory, offer the potential for overcoming some of the sampling and coding limitations of traditional content analysis" (p. 38). Riffe, Fico and Lacy (2014) designate a few categories for simple automated content analysis tasks: word counts, keyword-in-context, concordances, dictionaries, language structure (i.e., natural language processing), and readability. While all of these tasks are useful to specific research questions, most stop short of the annotation of data in a way that can directly test hypotheses. For example, counting the occurrence of words in a text (e.g., word counts), or the words around it (e.g., keywords in contexts and concordances) may be useful to the researcher to understand large chunks of data, but it is often only a first step in developing a scope for a content analysis (Conway, 2006). Similarly, while the way a particular sentence is written (e.g., language structure) and how readable it is (i.e., readability) are excellent annotations to have for data, they only offer a very narrow-scope of evidence that supports an even more narrow set of hypotheses (i.e. assumptions as to how well/poorly that passage was written).

Dictionary-based text analysis. Of the modern computer-assisted approaches, the dictionary-based approach is the most exhaustive content analysis method that computers can hope to automate for researchers (Riffe et al., 2014). It can not only provide context to data as the other methods can, but it can also be used to automatically classify text of any kind into groups of any kind. In fact, it is the most widely used approach in computer-assisted content analysis (West, 2001).

First attempted in 1968 at Harvard, the computerized-dictionary task is straightforward (Stone, Dunphy, Smith, & Ogilvie, 1968). Researchers assign lists of keywords that correspond to groupings (e.g., topics, attributes, or stakeholders) that they wish to identify in the text. The computer then scans each unit of analysis (e.g., a sentence or a paragraph) for the presence of those words. If a word from a list is present, then the computer annotates that unit as containing that grouping. Since then, this basic idea has been recreated with varying complexity and nuances in many different computer programs (see Riffe et al., 2014, p.170 for a review of these programs). Still, at the heart of these programs lie lists of words that researchers *manually* develop to represent constructs that they hope to identify.

Compared with the traditional manual content analysis, the dictionary-based approach increases the efficiency of text classification tasks to a great extent. A good number of recent journalism and communication studies have employed this method to analyze big social data for testing communication theories such as agenda setting and selective exposure (e.g., Neuman, Guggenheim, Jang, & Bae, 2014; Vargo, Guo, McCombs, & Shaw, 2014)

Dictionary-based text analysis still requires several subjective steps to adapt the content to the computer program. Like manual content analysis, the researcher needs to develop a predetermined list of categories as well as wordlists to indicate the categories. It is important to assess whether the predetermined list can adequately reflect the entire big dataset. In the past, reading a subset of news articles to discover the most covered topics or attributes for analysis was a reasonable approach for data of smaller sample size. However, with a dataset of one million or more units, researchers cannot even begin

to read a representative sample. Therefore, it is very likely that the predetermined list of categories will narrow or bias the potential areas to be analyzed. For these use cases where the categories that the researcher wishes to study are unknown initially, unsupervised machine learning methods may offer insight.

Unsupervised machine learning algorithms. In essence, unsupervised machine learning algorithms attempt to learn "hidden structure" in unlabeled data. One of the most popular approaches to do this involves topic modeling. The algorithm attempts to decompose data into contributions from multiple latent "causes" (topics) that are shared by all the data, but to different extents. To elaborate, a topic model views each document as an unordered "bag of words" which occur with different frequencies. It then "explains" the observed word frequencies in a given document in terms of a suitably weighted mixture of topical word frequencies where the weights indicate the different proportions of topics that appear in the document (Manning, Raghavan, & Schütze, 2009). For example, if an article contains the following words "gene", "dna", "rna", "evolve", "mutation", "data", "computational" and "statistics" in different proportions, then a topic model will view this article as a *mixture* of topics such as "genetics" (words such as "gene", "dna", "rna"), "evolution" (words such as "gene", "evolve", "mutation", "statistics"), and "data science" (words such as "data", "computational", "statistics") with the different proportions of words reflecting the article's topical emphasis. This approach is in fact ideally suited for processing text data in the context of journalism and mass communication because a document (e.g., a blog post or a tweet) is very likely to contain more than one topic. To discover the set of latent topics, many estimation and inference algorithms make use of the co-occurrence of words across documents. In our study, we

rely on the most widely used topic model, the Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003).

Despite the popularity of the LDA model in automating text analysis tasks, it has not yet been applied to answer research questions in the context of journalism and mass communication. Even in the field of computer science and engineering, the LDA algorithm has been most commonly applied to the analysis of well-constructed text documents such as newspaper and academic journal articles which are reviewed, edited, and proof-checked for grammar and spelling. Text data from social media in contrast presents a number of challenges. For instance, tweets are constrained to be short pieces of text (no longer than 140 characters) and are often terse, truncated, and quite "messy" because they contain abbreviations, symbols, and intentionally truncated words in addition to spelling errors and poor grammar. It has been recognized in the recent literature that automated topic modeling algorithms such as LDA, which work very well with well-constructed data, may fail to produce meaningful topics if applied to tweets directly without suitable preprocessing (Hong & Davison, 2010; Tang, Zhang, & Mei, 2013).

In this study we apply the LDA model with a suitable pre-processing step (described later) for data augmentation to discover prominent topics in Twitter's conversation about two political candidates, Barack Obama and Mitt Romney, during the 2012 U.S. presidential election. For the purpose of comparison, the dictionary-based approach was used to examine the same dataset. Specifically, the following research questions are studied.

R1a-b: Using a dictionary-based method, what is the qualitative structure and proportion of the topics in Twitter's coverage of Obama (a) and Romney (b) during the 2012 U.S. presidential election?

R2a-b: Using the LDA method with suitable pre-processing, what is the qualitative structure and proportion of the topics in Twitter's coverage of Obama (a) and Romney (b) during the 2012 U.S. presidential election?

R3: How do results generated by the two methods differ qualitatively and quantitatively?

Validity, Reliability and Cost

We cannot endorse any method of analysis without rigorous testing and evaluation. In social science, the three most important principles to evaluate any content analysis project including computer-assisted one are validity, reliability and cost (Riffe, Fico, & Lacy, 2014).

To determine how well the computer-generated results represent the actual meaning of the text, it is crucial for researchers to check the validity of the measurements. As Zamith and Lewis suggest (2015), "algorithms and dictionaries must often be repeatedly revised and tweaked to improve their performance" (p.4). The iterative process only concludes when the analysis yields a satisfactory level of construct validity. This is assessed when the researcher evaluates the algorithmic coder against the same coding decisions of a human and the two agree at an acceptable level.

When it comes to reliability, like manual content analysis, any human decision in the process of computer-assisted analysis must undergo intercoder reliability testing.

Fortunately, reliability is not a concern for the part of computer automation—assuming the algorithm implementation is correct—as computers are persistent and consistent.

Despite the importance of measuring reliability and validity in content analysis, few empirical big social data research—either in computer or social science—has explicitly elaborated on the process. As such, to what degree we can rely on these computer-generated results to answer research questions or test hypotheses remains suspicious. To address this concern, the study strictly follows the iterative process suggested by Zamith and Lewis (2015) in conducting the two analyses. Based on the computer-generated results, we further seek to explore which method can better capture the actual meaning intended in the text. Specifically, we ask:

R4: Which method, namely dictionary-based text analysis or LDA-based approach, produces more valid results?

Lastly, researchers should also compare the performance of different computational methods with respect to the cost involved in the analysis: the cost of human labor (e.g., time, expertise and knowledge of subject matter) as well as computational cost (e.g., execution speed, memory). This will be briefly discussed at the end of the paper, though a systematic comparison is beyond the scope of this study.

Method

For this methodological exploration, we collected data from Twitter during the 2012 U.S. presidential election. The Twitter API was called to retrieve relevant tweets during the sampled period. Specifically, the Streaming API call was used to retrieve public tweets from Twitter that mentioned the terms “Obama” or “Romney.” The collection started on

August 1st, 2012 and ended on November 28th, 2012, three weeks after the Election Day. In all approximately 77 million tweets were retrieved and stored in a 22-gigabyte corpus.

To examine how Twitter users discussed Obama and Romney separately, we then divided the original dataset into tweets that mentioned "Obama" but not "Romney," and tweets that mentioned "Romney" but not "Obama." For each group (i.e., Obama-only tweets and Romney-only tweets), we further filtered the dataset by only including Twitter users who posted at least four times (at least once a month) and at most 118 times during the sampled period (at most once a day). In other words, we only examined the most representative Twitter users by excluding those who were relatively inactive and those who were extremely active (e.g., robots) during the election. In total, 17,111,187 tweets about Obama authored by 1,599,918 unique Twitter users, and 11,022,299 tweets about Romney authored by 986,816 unique users were included in the final analysis. These two groups of tweets were used to test the two methods of computer-assisted text analysis discussed in this paper.

Method 1: Dictionary-based Issue Discovery

Following traditional journalism and mass communication research, this study used both deduction and induction to generate a comprehensive list of topic categories and the corresponding keyword index for analysis. We started with a literature review of the existing communication studies that examined news coverage and public opinion during U.S. political elections (e.g., Neuman et al., 2014; Vargo et al., 2014). The issue categories used in these studies formed the basis for our analysis.

Next, we conducted a preliminary analysis of the Twitter data in order to refine the issue categories identified earlier in the literature. However, with over 28 million

tweets to analyze, even reading one percent (n = 280,000) was far beyond the capacity of the researchers. Instead, as did Conway (2006) we began our analysis by taking a look at the most common words in the dataset. The entire corpus of tweets was stemmed, a process of reducing inflected or derived words to their word stem, base or root form (e.g., car, cars, car's, cars' → car). All the punctuations, numbers, extra spaces, special characters and stop words were also removed. Then, a term-frequency list was generated and sorted into a descending order. We then examined all words that occurred more than 1,000 times. Here, the list of issue categories derived from the literature review was adjusted based on the term-frequency results. The top words that the researchers thought corresponded directly to the adjusted issue categories were then placed into the wordlists for each issue.

Then, several rounds of reliability tests were performed to ensure the keywords lists were externally valid. A stratified sample of 1,800 tweets was pulled by each tweet group (i.e., Obama- and Romney-only tweets). Two human coders assigned each tweet by the issue it mentioned. Initial agreement for was 94%. Previous studies (e.g., Lombard, Snyder-Duch, & Bracken, 2002) consider 90% agreement a reasonable cutoff for robust intercoder reliability. Words that were correlated positively with the coders' annotations were then also placed into corresponding issue lists. A total of 16 issues were found through this analysis: (1) Tax; (2) Jobs/unemployment; (3) Federal budget deficit; (4) Economy in general; (5) Foreign affairs; (6) Immigration; (7) Health care; (8) Public order; (9) LGBT/same-sex marriage; (10) Abortion; (11) Environment/climate; (12) Energy; (13) Education; (14) Role of government; (15) Middle class; (16) Welfare.

Again, manual content analysis was performed to compare the human and computational results. A stratified sample of 800 tweets was then pulled for each of the 16 issues. Three rounds of reviewing the pulled samples and adjusting the word lists were performed. To achieve these results, two lists were established for identifying issues. Exact matching and non-exact matching lexicons were used to reduce false positive detection (e.g., "gas" returning matches for "Vegas"). The final agreement human to computer coding agreement was 97%. No issue scored below 92%. Intercoder reliability between humans was 100%. See Appendix A for a list of 16 issues and the associated keywords.

The keyword lists were then applied to each unit of analysis (i.e. a tweet) to detect whether any of the 16 issues were mentioned. Each issue was afforded a column and arranged in a rectangular data format so a unit could be coded as having any/all of the 16 issues. By tallying the number of occurrences of these 16 issues across the datasets, the topic proportions were calculated for Obama- and Romney-only tweets, respectively.

Method 2: Unsupervised LDA modeling

The unit of analysis in unsupervised LDA-based topic modeling is called a "document." The size of each document needs to be reasonably large for the LDA algorithm to extract meaningful topics. Using a single tweet with at most 140 characters as a document would produce misleading results due to its small size. To tackle the problem, one option is to combine a certain number of tweets based on some common features shared by these tweets such as authorship or time of posting. One study (Hong & Davison, 2010) chose to aggregate *all* tweets generated by the same author (across time) into a single document while another combined *all* tweets generated in certain unit of time (across all users) into

a single document (Zhao et al., 2011). The former approach mixes-up all topics across time for each user whereas the latter mixes-up topics across all users at each time unit.

In contrast to these two extremes, we propose a simple approach to combine tweets that preserves both time- and user-resolution of topics. Specifically, we chose to combine every four consecutive tweets from the same user into one document. In doing so, the Obama-only tweet dataset included a total of 4,820,018 documents. The Romney-only dataset included 3,091,220 documents.

In preparation for the LDA analysis, we further "cleaned" the datasets by stemming all the words and removing all the punctuations, spaces, numbers, special characters (e.g., hashtags, emojis, urls) and stop words. For each group of tweets (i.e., Obama-only and Romney-only), the remaining words were used to create a Document Term Matrix in which each row indicates a document and each column represents a word.

A Python package "Gensim" (Řehůřek & Sojka, 2010) was then used to train LDA over each group of tweets. In LDA modeling, the number of topics to be trained is at the discretion of the researcher. In our project, we decided the number of topics as 16 in the hope that the results here are comparable to those generated by the dictionary-based analysis. We adopted other parameters as suggested by Řehůřek and Sojka (2010).

The LDA training generated a list of 16 "topics" and probabilities of all the words associated to each topic. To determine what these "topics" actually meant, for each topic two communication researchers read all the corresponding words whose probability was higher than 1% and suggested a label that they felt represented the topic.

Finally, the proportion of the 16 topics in Obama-only and Romney-only tweets was calculated. The LDA algorithm also estimates the proportion score ("theta" in Blei et al., 2003) of each of the 16 topics in each document. In other words, the weight of each topic in each document was calculated (e.g., topic#1 – 60%, topic#2 – 20%, topic#3 – 20%). To calculate the proportion of a topic in each dataset, the theta value of each topic across the documents was tallied and then divided by the total number of documents.

Comparison and Human Evaluation

The results generated by the two computer-assisted text analysis methods were compared to each other. To explore the external validity of each method, the two sets of computer-generated results were then compared with human evaluations. A sample of 100 documents was pulled from Obama-only and Romney-only dataset, respectively. In each set of 100 documents, 32 documents were randomly selected to represent keyword-based results. In other words, at least two documents were labeled as containing each of the 16 predetermined topics. Likewise, 32 documents were randomly selected to represent LDA-based results. That is, in at least two documents each of the 16 LDA-generated topics was dominant (i.e., theta > 30%). The remaining 36 tweets were randomly selected from the entire dataset.

Two new independent researchers who had not previously seen the data read the documents and decided the topic for each document separately. They then discussed their coding results and recoded the data until they reached a 100% agreement. The human coding results were used to compare with the results generated by the two computer-assisted methods. For each document, they then decided which method generated "topic(s)" that were closer to their own coding results.

Results

R1 asked about the topic proportion in Twitter's coverage of Obama and Romney, respectively during the 2012 U.S. presidential election using the dictionary-based analysis. By using this method, the results show that the majority of the tweets in the sample did not mention any of the 16 topics identified earlier by the researchers. Specifically, out of 17,111,187 tweets mentioning "Obama," 85.7% of them did not specify any predetermined topic. Likewise, out of 11,022,299 tweets about "Romney," 84.7% of them contained no topic.

For tweets that did discuss at least one of the 16 pre-defined topics, the results of topic proportion in each dataset (i.e., Obama-only and Romney-only) are presented in Figure 1. The analysis found that "foreign affairs" (7.43%) was the most salient topic in Twitter's discussion about Obama, followed by "jobs/unemployment" (1.66%) and "health care" (1.63%). When it came to the conversation about Romney on Twitter, "tax" (4.80%) and "foreign affairs" (4.24%) were the two most discussed issues. On the other hand, certain topics such as "role of government" were only represented in a very small proportion of tweets about either political candidate.

<Insert Figure 1 about here>

In answering R2, the LDA-based analysis automatically discovered 16 most important "topics" inherent in Twitter's coverage for Obama and Romney, respectively. Table 1 presents the 16 LDA-generated topics about Obama, each followed by a list of words ranked according to their probability of relevance to each topic. The label of each topic category and the percentage of tweets that contained each topic are also detailed in

Table 1. In the same format, Table 2 presents the topic information of tweets mentioning Romney.

<Insert Table 1 and Table 2 about here>

In the case of tweets mentioning Obama, the researchers were not able to determine a coherent, sensible issue for topic#4 or topic#15 (see Table 1). In combination, this represented a substantial proportion of tweets (35.75%). For the rest of the tweets, the most salient topic (17.86%) was about the presidential campaign and election. "Foreign affairs," specifically the Benghazi attack (14.73%), was the second most prominent topic in Twitter's discussion about Obama. It is also worthwhile to note that one of the most important topics about Obama on Twitter was in Spanish (2.48%), a pattern not found in Twitter's conversation about Romney.

Table 2 illustrates the LDA-generated topics regarding Romney. Like Twitter's coverage of Obama, nearly half of the tweets (44%) that mentioned "Romney" did not correspond to any meaningful topic. The other 15 LDA-generated topics appeared to be somewhat evenly distributed among the tweets. Notably, uncivil discourse (6.40%) was the most prominent "topic" in tweets about Romney. The percentage of tweets about Romney that contained this topic (6.40%) was twice as much as that about Obama (3.12%). Campaign/election & Education (5.96%) and the impact of tax cuts on middle class (4.55%) were also salient in tweets about Romney.

R3 sought to compare the results generated by two research methods. It appeared that many of the LDA-generated "topics" overlapped with those identified by the researchers. Remarkably, both approaches discovered that "foreign affairs" was a salient topic in Twitter's conversation about Obama, and that taxation was closely associated

with the discussion about Romney. In some regard, the two research methods were similar.

On the other hand, notable differences emerged. Though neither automated method could determine the content of *all* tweets, the results show that the LDA-based analysis was able to infer topics for more tweets than the dictionary-based approach. This is likely because the latter relied on a list of pre-determined topics with a very limited number of keywords (see Appendix A). If a tweet did not contain a keyword, it was dismissed. On the other hand, without a predetermined "codebook," the LDA-based analysis discovered a wider variety of topics. Some of these topics were not identified by the pre-existing research on elections or the researchers earlier in the project. For example, Romney's airplane window gaffe[1] went "viral" on Twitter. This topic was captured in the LDA analysis (topic#2, 2.8%) but not by the dictionary-based method. While both methods found "foreign affairs" to be a salient topic in Twitter's coverage of Romney, the LDA analysis revealed more detail. According to Table 2, 4.08% of the tweets were about the Benghazi attack in Libya (topic#15) and 3.14% referred to the U.S. trade with China (topic#16).

In turn, certain issues that were considered important by the researchers and thus included in the dictionary-based analysis turned out to be absent in the LDA-generated results. For example, the dictionary-based analysis shows that about 0.15% of the tweets mentioning either candidate discussed the immigration issue, a subject that was not captured by the LDA-based approach. However, it should be noted that in this study LDA

[1] During a fundraiser in California, Romney referred to an emergency landing by an airplane carrying his wife and said that "when you have a fire in an aircraft, there's no place to go, exactly, there's no – and you can't find any oxygen from outside the aircraft to get in the aircraft, because the windows don't open." Critics seized on his lighthearted objection to windows that "don't open" as another example of his outlandish offensives against common sense.

was forced to discover only 16 topics. Had LDA been run with 20 topics, for example, it might have been able to capture less salient topic such as immigration as well.

What also differentiates the two methods is that while each of the 16 predetermined topics for the dictionary-based analysis is distinct and includes one subject, one LDA-generated "topic" may contain multiple themes. Consider the tweets about Obama. Topic#6 referred to both foreign affairs and health care, and topic#7 contained three issues: federal budget, tax and Medicare. These LDA-generated "topics" with mixed information provide insights into how Twitter users associated different subjects in talking about the given candidate.

R4 asked about the external validity of the machine coding results. Two communication researchers read a sample of 100 documents about each politician and then compared their decisions with those generated by the two computer-assisted text analysis approaches. As Table 3 demonstrates, each method failed to interpret a considerable number of documents as human coders did. For 21 documents about Obama and 31 about Romney, neither method captured what the text actually meant.

Despite the misinterpretations, the results show that the performance of the LDA-based analysis was better than that of the dictionary-based analysis according to human evaluations. Out of 92 documents mentioning Obama that were decipherable by human coders, the LDA-based analysis succeeded in capturing the main idea of half of them (N=58), whereas the dictionary-based approach captured slightly over a third (N=35). Likewise, for the 99 readable documents mentioning Romney, the LDA-based analysis aligned with the human coding in 57 documents, while the dictionary-based approach made correction decisions for only 23 documents.

The difference may be explained by the fact that a great number of tweets contained content that was beyond the limited vocabulary on which the dictionary-based analysis relied. Consider the following document as an example:

> #obama http://t.co/srx5hyvd rt @campaignsosa300: love sosa remix!! #obama #300 http://t.co/cxxxjktv rt @followmeobama: rosa parks sat, so martin luther king jr. could walk, so barack obama could run, so we can all fly... r-t to show lov.

The LDA analysis determined that two salient topics inherent in this document were "#5 Martin Luther King & African American civil rights" (49%) and "#11 Campaign/election" (11%). However, this is not a topic identified earlier for the dictionary-based analysis.

While the dictionary-based analysis failed to discover certain topics (i.e., a false negative), the main validity concern regarding the LDA-based approach comes from its false detection (i.e., a false positive). The following document provides an example.

> rt @donaldjtrumpjr: love that someone criticizing me said that half the 16 trillion was inherited by obama. moron the other 43 president ... @michelleobama not for obama!

Human coders found that this document is about federal budget deficit. The dictionary-based analysis improperly indicated no presence of this topic. On the other hand, the LDA-based analysis improperly indicated that "#9 foreign affairs" represented 14% of the content of this document.

Indeed, Twitter users used sarcasm quite often when referring to the two political candidates, which in fact would be a challenge to any computer-assisted text analysis not using advanced natural language processing. Here is an example:

> rt @djbigwill: you can pay attention to the words romney says, but watch his mannerisms... it tells the story he's too scared to say. rt @garyowencomedy: so romney is mexican? next thing he's gonna say is "i had an abortion it was horrible" rt @chrisrockoz: mitt romney is like a best buy employee trying to sell you something he cannot fully explain. #debate #debate'

Human coders agreed that this document referred to Romney's credibility. The expressions that "romney is Mexican" and "i had an abortion" were rhetoric techniques to emphasize that Romney's behaviors were not consistent with what he said. However, while the dictionary-based analysis determined the document was about "abortion," the LDA-based analysis determined that 10% of the content of this document was about "#9 foreign affairs." Both decisions produced misleading information for the final analysis of topic proportion.

Discussion

This paper presents an empirical study that investigated and compared two computer-assisted text analysis methods: (1) the dictionary-based analysis, perhaps the most popular automated analysis approaches in social science research; and (2) unsupervised topic modeling (i.e., LDA analysis), one of the widely used algorithms in the field of computer science and engineering. By applying two different "big data" methods to make sense of the same dataset—77 million tweets about the 2012 U.S. presidential election, the study provides a starting point for scholars to evaluate the efficacy and validity of different computer-assisted methods for conducting journalism and mass communication research, especially in the area of political communication.

Overall, the study suggests that both computer-assisted text analysis methods generated some valuable information from the big dataset. According to both approaches, Twitter users were most likely to mention foreign affairs in their discussion of Obama and they tended to relate Romney with the issue of taxation. These kinds of summary statistics can be compared to media coverage of both candidates to examine journalism and communication theories such as media effects and issue ownership.

More importantly, the research found that the two approaches differed to a large extent in the results they produced. The LDA-based analysis performed better than the dictionary-based approach in several aspects. Specifically, the LDA-based analysis was able to interpret more tweets and reveal more nuanced details of the conversation. Based on the evaluations of human coders, the LDA-based analysis was also found to be more valid than the dictionary-based approach. Given that the LDA-based analysis involves only a minimum amount of human labor, it is in fact also more cost-effective than other computer-assisted methods. It should be highlighted that, as far as we know, this present study is the first attempt to apply the LDA analysis to the context of journalism and mass communication research. Considering its superior performance, future research should consider using this method to analyze text data in other subject areas.

The dictionary approach does however still maintain a few uses cases. When researchers are only looking at a specific issue or topic, building an issue list may be easier than analyzing an entire corpus. Moreover, the dictionary approach remains more "focused."

However, this study also clearly demonstrates that significant errors were found in results generated by *both methods*.

- LDA yielded more false positives.
- The dictionary-based approach produced more false negatives.

This is a challenge presented by the task of deciphering messages on Twitter. Tweets are terse, extremely unstructured and often involve sarcastic expressions. Journalism and mass communication scholars should make notes of these potential misinterpretations in their big data studies. Future research should also consider combining the two methods. For instance, if a researcher had an initial list of issues in mind to study, it would not hurt to consult the literature for wordlists associated with that issue. However, it could be more advantageous to take those wordlists and compare them to the populous LDA topics in the corpus. Initial lists of words could then be augmented to include those words. Conversely, LDA topics can be used to "induce" popular issues or topics. They are, however, messy. From here, researchers can remove erroneous words and "clean up the lists" to make them more externally valid.

Beyond this, other advanced machine learning algorithms can reveal even more meaning from Twitter data. For example, there appears to be some recent progress in using natural language processing and topic modeling methods to detect properties of language such as negation and sarcasm (Rajadesingan, Zafarani, & Liu, 2015).

Valuable social science lessons should be learned from "the algorithmic coder." Most importantly, journalism and mass communication scholars must understand the "side effects" of the new methodological choices they face. Here we provide suggestions based on different use cases and research aims. Understanding why certain algorithms perform better than others is crucial to externally valid results. The "secret sauce" or the gears behind algorithms are almost always encoded in peer reviewed social science

research. Journal articles themselves are not good repositories for computer scripts (e.g., python code) that these algorithms are created with. Beyond this, big data sets remain hard to share and thus replication becomes a major issue. Without transparent methodological descriptions, the majority of big data social science work may not validly measure constructs in text. As such, emerging methods in computational science need to be investigated with substantial rigor to determine whether they are externally valid enough to measure the construct in which they are intended. The current study presents an example of such an endeavor and shows clear pros and cons.

Reference

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Connolly-Ahern, C., Ahern, L. A., & Bortree, D. S. (2009). The effectiveness of stratified constructed week sampling for content analysis of electronic news source archives: AP Newswire, Business Wire, and PR Newswire. *Journalism & Mass Communication Quarterly*, *86*(4), 862–883.

Conway, M. (2006). The subjective precision of computers: A methodological comparison with human coding in content analysis. *Journalism & Mass Communication Quarterly*, *83*(1), 186–200.

Goel, S., Anderson, A., Hofman, J., & Watts, D. (2013). The structural virality of online diffusion. Journal of Management Science. *Journal of Management Science*. Retrieved from http://www.jakehofman.com/inprint/twiral.pdf

Hester, J. B., & Dougall, E. (2007). The efficiency of constructed week sampling for content analysis of online news. *Journalism & Mass Communication Quarterly*, *84*(4), 811–824.

Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics* (pp. 80–88).

Kamhawi, R., & Weaver, D. (2003). Mass communication research trends from 1980 to 1999. *Journalism & Mass Communication Quarterly*, *80*(1), 7–27.

Krippendorff, K. (2004). Reliability in content analysis. *Human Communication Research*, *30*(3), 411–433.

Lacy, S., Duffy, M., Riffe, D., Thorson, E., & Fleming, K. (2010). Citizen journalism web sites complement newspapers. *Newspaper Research Journal*, *31*(2), 34–46.

Leccese, M. (2009). Online information sources of political blogs. *Journalism & Mass Communication Quarterly*, *86*(3), 578–593.

Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content analysis in an era of big data: A hybrid approach to computational and manual methods. *Journal of Broadcasting & Electronic Media*, *57*(1), 34–52.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, *28*(4), 587–604.

Luke, D. A., Caburnay, C. A., & Cohen, E. L. (2011). How much is enough? New recommendations for using constructed week sampling in newspaper content analysis of health stories. *Communication Methods and Measures*, *5*(1), 76–91.

Manning, C., Raghavan, P., & Schütze, H. (2009). Flat clustering. In *Introduction to Information Retrieval* (pp. 349–374). New York: Cambridge University Press.

McQuail, D. (1994). *Mass communication theory: An introduction* (3rd ed). Thousand Oaks, CA: Sage.

Morales, L. (2012, September 21). U.S. Distrust in Media Hits New High. Retrieved from http://www.gallup.com/poll/157589/distrust-media-hits-new-high.aspx

Napoli, P. M. (2011). *Audience evolution: New technologies and the transformation of media audiences*. New York: Columbia University Press.

Neuman, R. W., Guggenheim, L., Jang, S. M., & Bae, S. Y. (2014). The dynamics of public attention: Agneda-setting theory meets big data. *Journal of Communication*, *64*, 193–214.

Rajadesingan, A., Zafarani, R., & Liu, H. (2015). Sarcasm detection on Twitter. *Web Search and Data Mining (WSDM)*.

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50).

Riffe, D., & Freitag, A. (1997). A content analysis of content analyses: Twenty-five years of Journalism Quarterly. *Journalism & Mass Communication Quarterly*, *74*(3), 515–524.

Riffe, D., Lacy, S., & Fico, F. (2014). *Analyzing media messages: Using quantitative content analysis in research*. New York: Routledge.

Rosen, J. (2008). A most useful definition of citizen journalism. *PressThink*.

Rosen, S. (2010). Is the Internet a Positive Force in the Development of Civil Society, a Public Sphere, and Democratization in China? *International Journal of Communication*, *4*, 509–516.

Stone, P., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1968). The general inquirer: A computer approach to content analysis. *Journal of Regional Science*, *8*(1), 113–116.

Tang, J., Zhang, M., & Mei, Q. (2013). One theme in all views: Modeling consensus topics in multiple contexts. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 5–13).

Vargo, C., Guo, L., McCombs, M., & Shaw, D. L. (2014). Network issue agendas on Twitter during the 2012 US presidential election. *Journal of Communication*, *64*(2), 296–316.

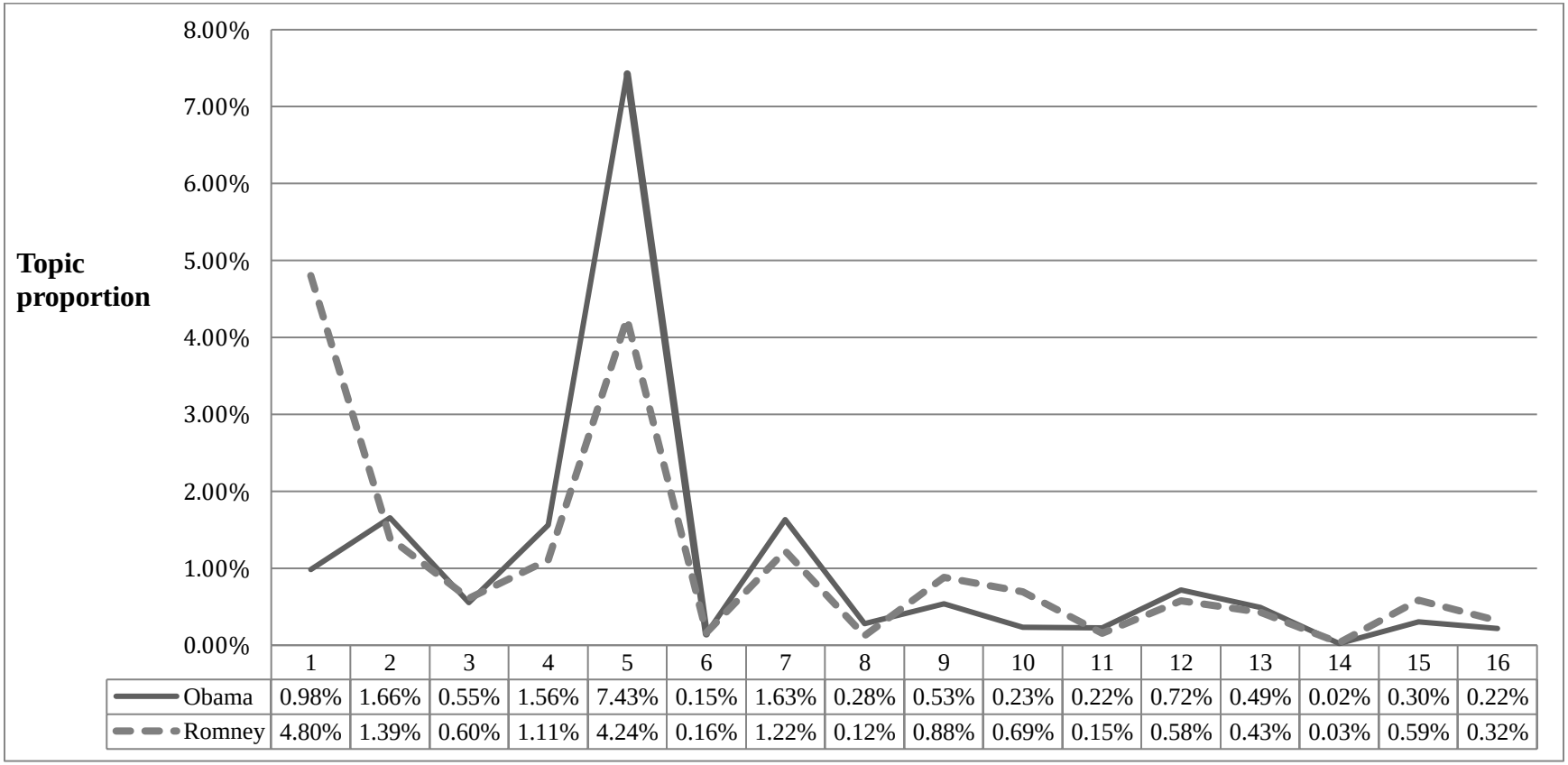
West, M. D. (Ed.). (2001). *Theory, method, and practice in computer content analysis* (Vol. 16). Greenwood Publishing Group.

Xiang, D. (2013). China's image on international English language social media. *Journal of International Communication*, *19*(2), 252–271.

Zamith, R., & Lewis, S. C. (2015). Content analysis and the algorithmic coder: What computational social science means for traditional modes of media analysis. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 307–318.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., & Li, X. (2011). Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval* (pp. 338–349). Berlin Heidelberg: Springer.

Figure 1. Topic proportion of Twitter's coverage of Obama and Romney (Dictionary-based analysis)



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Obama	0.98%	1.66%	0.55%	1.56%	7.43%	0.15%	1.63%	0.28%	0.53%	0.23%	0.22%	0.72%	0.49%	0.02%	0.30%	0.22%
Romney	4.80%	1.39%	0.60%	1.11%	4.24%	0.16%	1.22%	0.12%	0.88%	0.69%	0.15%	0.58%	0.43%	0.03%	0.59%	0.32%

Notes:
1. The sample includes a total of 17,111,187 tweets about Obama and 11,022,299 tweets about Romney.
2. Topics: (1) Tax; (2) Jobs/unemployment; (3) Federal budget deficit; (4) Economy in general; (5) Foreign affairs; (6) Immigration; (7) Health care; (8) Public order; (9) LGBT/same-sex marriage; (10) Abortion; (11) Environment/climate; (12) Energy; (13) Education; (14) Role of government; (15) Middle class; (16) Welfare.

Table 1: Top 16 “topics” on Twitter’s coverage of Obama

	“Topic” and associated words	Label	Proportion (N=17,111,187)
1	year, bush, blame, four, anoth, done, problem, ago, yr, georg, fix, food, destroy, economi, mess, hasn, stamp, welfar, fault, clean, realiz, longer	Blamed former President George W. Bush for the economic turmoil that has beset the United States in the past few years	2.88%
2	news, clinton, sandi, new, bill, video, fox, hurricane, hit, photo, christi, york, chris, storm, hillari, fundrais, declar, star, prais, reddit, jersey, sent, song, isaac, katrina	Hurricane Sandy	3.12%
3	vote, regist, count, black, go, tomorrow, everybodi, teamobama, pleas, moreyear	Campaign/election (e.g., voting for Obama)	3.67%
4	say, like, just, get, peopl, go, don, now, want, know, make, one, michell, think, need, right, time, see, said, look	N/A	33.40%
5	mean, famili, money, left, behind, set, street, letter, walk, dream, car, king, accept, sticker, nobodi, fli, park, martin, luther, pack	Martin Luther King and African American civil rights	1.58%
6	kill, care, bin, laden, health, home, gave, els, decis, iraq, osama, war, troop, politician, healthcar, end, got, back, seal, want, women, better	Foreign affairs (i.e., withdrawal of troops, Iraq, bin Laden) & Health care	2.01%
7	tax, fact, night, cut, offic, busi, debt, pay, took, check, medicar, rnc, eastwood, spend, unemploy, clint, last, rais, deficit, small, budget, trillion, plan, increas, chair, billion, build, sinc, gop, spent, job, taken	Federal budget & Tax & Medicare	2.77%
8	speech, dnc, watch, ve, forward, speak, move, convent, turn, chang, michell, differ, leav, democrat, minut, seen, togeth, live, nation	The Democratic National Convention & Michelle Obama	3.87%
9	american, tcot, benghazi, gop, new, attack, lie, video, endors, ad, term, polici, libya, administr, report, media, plan, call, record	Foreign affairs (i.e., the Benghazi attack)	14.73%
10	job, million, women, middl, creat, jay, class, fight, rate, student, powel, beyonc, dollar, educ, colin, auto, industri, grow, unemploy, colleg, afford, eat, economi, loan, dinner, wtf, pay, equal, save	Jobs & Social groups (e.g., women, students and middle class) & Celebrities (i.e., Fundraising reception hosted by Jay-Z and Beyoncé)	2.63%
11	elect, will, win, debat, campaign, day, poll, state, today, republican, voter, back, work, america, won, ohio, show, big, lead	Campaign/election (e.g., Obama will win or won the election)	17.86%
12	fuck, shit, nigga, im, ya, win, gone, bitch, ain, yo, yall, gotta, ass, don't, lose, bout, mama, europ, move, wit, aint, dat, lmao, daddi, grandma, pin	Incivility	3.12%
13	white, hous, post, old, approv, enough, folk, pic, ballot, pictur, shirt, birth, wear, burn, due, member, facebook, rock, certif, black	Campaign/election (e.g., Posting ballot picture)	1.87%
14	usa, para, president, su, las, di, del, es, va, son, est, ha, da, estado, per, una, dan, et, lo,	[Spanish]	2.48%

	pa, il, non, si, les, como, mi, ma, qu, ,vega, washington, latino, dem		
15	pleas, yes, follow, god, friend, mr, retweet, kid, happi, yeah, dear, birthday, direct, add, boy, parent, brother, candi, teacher, halloween, justin, name	N/A	2.35%
16	use, never, question, isn, gay, answer, surpris, pro, gas, defend, abort, marriag, denver, latino, shot, price, legal, offic, slogan, campaign, marijuana, fire	Individual liberties (e.g., LGBT, abortion) & Energy & Drug issues	1.66%

Note:
For each topic, all words with probability larger than 1% were included in the list. The words were ranked based on the probability estimated by the LDA model. The words were stemmed.

Table 2: Top 16 “topics” on Twitter’s coverage of Romney

	“Topic” and associated words	Label	Proportion (N=11,022,299)
1	tax, pay, plan, return, cut, middl, class, releas, paid, rais, million, medicar, rate, budget, year, reid, incom, dollar, welfar	Tax & Middle class	4.55%
2	look, like, away, take, bill, open, pass, suck, walk, window, babi, dick, result, spent, clinton, roll, marriag, halloween, test, star, airplan, type, porn, cri, plane, asshol	Romney’s airplane window gaffe	2.80%
3	women, care, gay, black, poor, peopl, woman, understand, rich, health, high, die, american, kind, dog, don, right, doesn, flip, children, equal, men, school, relat, decis, mexican, flop, approv, presid, sick, view	Social groups (e.g., women, LGBT, the poor) & Health care	3.53%
4	tcot, pick, vp, campaign, romneyryan, ralli, run, ohio, state, gop, mate, announc, endors, event, choic, crowd, donat, unit, florida, will, ticket, vice, red, va, today, wisconsin, rep, pa, join, bus	Campaign/election (e.g., Romney chose Ryan as vice presidential running mate)	4.01%
5	slogan, american, colleg, parent, money, bitch, mom, said, keep, america, use, moment, kkk, get, even, realiz, borrow, talk, make, fuck, bag, african, commun, small, know, akward, alreadi, full, ur, akward, best, pay, chip, ass, wasn, nigga	Romney’s campaign slogan & Education	2.72%
6	poof, sandi, car, fema, win, go, game, hurrican, relief, disast, cotton, back, took, field, da, storm, crib, volunt, hunger, feder, mil, tribut, teaparti, hunnit, elect, victim, mit, detroit, shut	Hurricane Sandy	1.89%
7	ann, speech, gop, rnc, convent, dnc, street, christi, eastwood, chris, republican, nomin, ron, offic, hors, sesam, wall, olymp, accept, quot, tampa, top, love, success	The Republican National Convention & Ann Romney	2.83%
8	video, full, comment, binder, women, post, secret, percent, machin, tagg, facebook, mother, blog, style, washington, new, fundrais, social, latino, tape, leak, campaign, remark, caught, american, page, israel, ohio, voter, servic	Presidential debate (e.g., Romney’s “binders full of women” comment)	3.46%
9	debat, poll, big, news, last, night, voter, lead, bird, cnn, fox, show, new, state, predict, ahead, victori, endors, presidenti	Presidential debate (Romney’s “big bird” comment) & Campaign/election (e.g., polling results, endorsement)	4.33%
10	say, just, like, will, get, think, go, want, don, one, peopl, make, know, now, said, debat, elect, see, right, campaign, support, time, need, realli, america, lie, tell, talk, let, good, call, won, take, even, man, back	N/A	44.93%
11	vote, win, elect, will, colleg, move, god,	Campaign/election & Education	5.96%

	world, go, student, aid, want, get, cut, tomorrow, chanc, countri, ha, futur, dear		
12	becom, make, around, gone, will, hoe, stamp, danc, fuck, food, term, struggl, band, condit, start, twerk, might, noodl, work, read, ramen, lil, name, know, flavor, even, weav, away, gon, don, nigga	Food stamps	2.69%
13	white, vote, hous, nicki, abort, minaj, govern, birth, peopl, want, racist, state, massachusett, republican, control, pro, shirt, wear, dash, rape, lose, don, stacey, endors, economi, lazi, certifi, black, support, leader	Individual liberties (e.g., birth control, abortion)	2.67%
14	fuck, win, shit, ass, vote, nigga, bitch, bet, im, get, ain, don't, lmao, teacher, kid, black, class, sticker, move, cuz, wanna, canada, smh	Incivility	6.40%
15	republican, polici, candid, question, attack, foreign, parti, presidenti, governor, issu, gop, answer, democrat, john, media, race, respons, libya, press, mccain, campaign, critic, akin, american, economi, senat, advis, death, palin, call, step, comment	Foreign affairs (e.g., Libya)	4.08%
16	job, bain, million, china, busi, creat, usa, compani, capit, auto, iran, worker, plan, debt, energi, gov, ad, new, economi, invest, bailout, profit, church, jeep, employe, american, outsourc, deal, ma, save, fix, mormon, govern	Job creation & Foreign affairs (e.g., China, Iran)	3.14%

Note:
For each topic, all words with probability larger than 1% were included in the list. The words were ranked based on the probability estimated by the LDA model. The words were stemmed.

Table 3. Comparison of marching and human coding

	Obama-only (N=100)	Romney-only (N=100)
Both methods captured the main idea of the document	22	12
Neither method captured the main idea of the document	21	31
Dictionary-based analysis captured the main idea of the document but LDA-based analysis did not.	13	11
LDA-based analysis captured the main idea of the document but dictionary-based analysis did not.	36	45
N/A (i.e., human coders cannot decipher the content)	8	1

Appendix A

Topic 1: Tax

- taxes = ["tax"]

Topic 2: Jobs/unemployment

- unemployment = ["employment", "employed"]
- unemploymentexact = ["jobs", "job growth", "job creation","lay off", "laid off", "out of work"]
- notunemployment = ["steve jobs"]

Topic 3: Federal budget

- fedbudget = ["deficit","budget"]
- fedbudgetexact = ["federal debt", "government debt", "national debt", "debt ceiling", "fiscal cliff", "spending cut", "government shutdown"]

Topic 4: Economy in general

- economy = ["economic", "recession"]
- economyexact = ["economy","recovery", "recoveries", "inflation","stock market", "dow", "GDP", "gross domestic product"]

Topic 5: Foreign affairs

- foreignaffairs = ["terrorist","foreign","Iraq","Iran", "Afghan", "Israel","Islam","Palestinian","Arab", "Syria", "Libya","troop","outsource", "insource","Russia"]
- foreignaffairsexact = ["Benghazi", "United Nation", "US embassy", "U.S. embassy", "Ahmadinejad", "Putin", "Chaves", "Castro", "Kim Jong-un", "North Korea", "North Korean", "world leaders", "nations", "hamas", "terrorism", "war on terror", "Osama", "bin Laden", "al Qaeda", "China", "Chinese", "trade", "cheap labor", "currency manipulation", "world trade organization", "middle east", "middle eastern", "Saddam", "Persian Gulf", "Muslim", "Palestine", "North Africa", "North African", "Asia", "overseas", "Taliban", "Yemen", "homeland security", "national security", "Pentagon", "military", "defense", "CIA", "armed forces"]

Topic 6: Immigration

- immigrationexact = ["immigration", "immigrant", "immigrate", "DREAM Act", "border issue", "border issues", "border safety","border security", "deportation"]

Topic 7: Health care

- healthcareexact = ["health", "healthcare", "medical", "Obamacare", "affordable care", "Romneycare", "Medicare", "Medicaid"]

Topic 8: Public order

- publicorderexact = ["illegal drug", "marijuana", "heroin", "cocaine", "methamphetamine", "drug trade", "drug addiction", "drug abuse", "alcoholism",

"alcohol addition", "alcohol abuse", "gun control", "gun rights", "firearm", "NRA", "crime rate", "prisons", "law enforcement", "death penalty"]

Topic 9: LGBT/same-sex marriage

- lgbtexact = ["don't ask, don't tell", "LGBT", "lesbian", "gay", "homosexual", "same-sex", "same sex"]

Topic 10: Abortion

- abortionexact = ["planned parenthood", "contraception", "abortion", "pro choice", "pro life", "Wade", "reproductive rights"]

Topic 11: Environment/climate

- environmentexact = ["renewable", "environmental", "pollution", "pollute", "pollutes", "clean air", "global warming", "climate change", "wildlife", "clean water", "natural resource", "sea levels", "sustainable development"]

Topic 12: Energy

- energyexact = ["gas", "energy", "oil", "coal", "drill", "drilling"]

Topic 13: Education

- educationexact = ["classroom", "education", "teachers", "tuition", "schools", "school voucher", "failing school", "school choice", "No Child Left Behind", "academic"]

Topic 14: Role of government

- governmentexact = ["nationalize", "nationalizes", "nationalized","nationalizing", "nationalization", "role of government", "size of government", "big government", "bigger government", "small government", "smaller government", "overbearing government", "government intervention"]

Topic 15: Middle class

- middleclassexact = ["middle class"]

Topic 16: Welfare

- welfareexact = ["welfare"]