Journalism & Mass Communication Quarterly
1–28

DOI: 10.1177/1077699016639231
jmcq.sagepub.com
SAGE

Big Social Data Analytics in Journalism and Mass Communication: Comparing Dictionary-Based Text Analysis and Unsupervised Topic Modeling

Lei Guo[1], Chris J. Vargo[2], Zixuan Pan[3], Weicong Ding[4], and Prakash Ishwar[1]

Abstract

This article 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., Latent Dirichlet Allocation [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

[1]Boston University, MA, USA
[2]The University of Alabama, Tuscaloosa, USA
[3]Yodlee, Redwood City, CA, USA
[4]Technicolor Research, Los Altos, CA, USA

Corresponding Author:
Lei Guo, Division of Emerging Media Studies, Boston University, 704 Commonwealth Ave. 304D, Boston, MA 02215, USA.
Email: guolei@bu.edu

In the field of journalism and mass communication, researchers seek to answer questions such as "How do the news media or other emerging media outlets cover an issue, a campaign, or a public figure?" "Do traditional news media still have significant effects on public opinion?" and "Can public conversations on social 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. With these new possibilities, however, come new challenges. The size of the data—for example, 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 article presents a methodological exploration of computer-assisted methods for processing big *social* data—text-based big data collected from various social networking sites. *Big data* is a broad term used for datasets that have a size (e.g., dimensionality, volume, and velocity of generation) and complexity (e.g., diversity, variability) that exceed the capabilities of *traditionally used tools* for capturing, processing, curating, and analyzing data within a tolerable timeframe (e.g., Beyer & Laney, 2012; Laney, 2001). What size qualifies as "big data" is domain-dependent and ever evolving. In social science, "big data" refers to datasets that are too big for humans to code a representative sample of the entire dataset (Riffe, Lacy, & Fico, 2014). In other words, the criterion to evaluate whether a dataset is "big" is a function of the amount of time required for a human to make a decision on a given unit. For complex problems and large documents, datasets that extend beyond 10,000 may be considered "big." For smaller documents that require little time per unit to code (e.g., tweet), datasets in the 100,000s are generally considered too large for manual methods (Riffe et al., 2014).

To process "big data" that is beyond the capabilities of manual analysis, 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 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 article hopes to provide a platform for scholars to assess which computer-assisted 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% 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 et al., 2014). As examples, a manual content analysis approach can answer research questions such as "What is the most salient topic in the news coverage of a political election?" and "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) and/or 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.

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. 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 articles suggest that sample size varies by subject domain, by media being sampled, and by variable type being analyzed. Although Riffe, Lacy, and Fico (2014) 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 (p. 93). As a result, no clear sampling guidelines exist for big datasets 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) are early to recognize the value of computer-assisted 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 et al. (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. Although 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, although 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 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 word lists 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 typical topic model views each document as an unordered "bag of words" which occurs with different frequencies.[1] 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 within each document. In our study, we rely on the most widely used topic model, the Latent Dirichlet Allocation (LDA; Blei, Ng, & Jordan, 2003). The LDA model assumes that the topic proportion of each document is sampled from a

Dirichlet distribution. We note that other bag-of-words-based topic models have been studied by assuming different distributions such as the log-normal distribution in Correlated Topic Model (Blei & Lafferty, 2006).

Many researchers in the field of computer science and engineering have used the LDA algorithm to examine journalism and mass communication text (e.g., Newman, Chemudugunta, Smyth, & Steyvers, 2006; Zhao et al., 2011). It is important to note that 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). A more recent development of topic model, which can potentially improve the interpretability of topics discovered, is to involve humans to iteratively refine the topics produced by an automated topic-modeling algorithm (Chuang et al., 2015; Hu, Boyd-Graber, Satinoff, & Smith, 2014). While promising, we leave such human "in-the-loop" approaches to future work. It should also be noted that, despite the popularity of the LDA model in automating text analysis tasks, relatively few journalism and mass communication researchers have applied the approach to answer their research questions.

In this study, we apply the LDA model with a suitable preprocessing 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.

RQ1a-RQ1b: 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?
RQ2a-RQ2b: Using the LDA method with suitable preprocessing, 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?
RQ3: 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 et al., 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 (2015) suggest, "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,

RQ4: Which method, namely dictionary-based text analysis or LDA-based approach, produces more valid results?

Finally, researchers should also compare the performance of different computer-assisted 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 article, 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 application programming interface (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 1, 2012, and ended on November 13, 2012. In all, approximately 77 million tweets were retrieved and stored in a 22-gigabyte corpus.

To examine research questions that asked 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."[2] 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 4 times (i.e., at least

once a month). In each group, an average user produced five tweets. In other words, we only examined relatively active Twitter users.[3] In total, 30,061,046 tweets about Obama authored by 1,631,095 unique Twitter users, and 18,677,277 tweets about Romney authored by 1,007,421 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 article.

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 thorough literature review of the existing communication studies that examined news coverage and public opinion during U.S. political elections, both previous ones (e.g., Kiousis, 2004; McCombs, 2014; Petrocik, 1996) and the 2012 presidential election (e.g., Neuman et al., 2014; Vargo et al., 2014). The issue categories used in these studies formed the basis for our analysis. In other words, we strictly followed the traditional approach in journalism and mass communication research in deciding initial issue categories.

Next, we conducted a preliminary analysis of the Twitter data to refine the issue categories identified earlier in the literature. However, with over 48 million tweets to analyze, even reading 1% (n = 480,000) 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 word lists for each issue.

Then, several rounds of reliability tests were performed to determine the final issue categories for the analysis and to ensure the keyword lists were externally valid. We first conducted an intercoder reliability test to examine the reliability of the issue categories. A stratified sample of 1,800 tweets was pulled by each tweet group (i.e., Obama-only and Romney-only tweets). Two human coders assigned each tweet by the issue categories it mentioned. Initial intercoder reliability in terms of percent agreement for this variable (i.e., issue category) was 94%. Previous studies (e.g., Lombard, Snyder-Duch, & Bracken, 2002) consider 90% agreement a reasonable cutoff for robust intercoder reliability. The two coders then discussed the discrepancies of their coding to refine the issue categories. A total of 16 issues were decided 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) lesbian, gay, bisexual, transgender (LGBT)/same-sex marriage; (10) abortion;

(11) environment/climate; (12) energy; (13) education; (14) role of government; (15) middle class; and (16) welfare. Words that were correlated positively with the coders' annotations were then also placed into corresponding issue lists.

Manual content analysis was then performed to refine the keyword lists. Because computers are automated systems that are persistent and reliable, measures of intercoder reliability do not need to be calculated for computer-coded data (Riffe et al., 2014; Zamith & Lewis, 2015). What is in question is how externally valid that result may be. Therefore, we conducted both inter-human-coder reliability tests as well as human-computer agreement tests. A stratified sample of 800 tweets was then pulled for each of the 16 issues. For each issue, the two human coders read each of the 800 sampled tweets and decided whether the given issue was mentioned (i.e., a binary coding). Intercoder reliability between humans was 100% for all the 16 issues. This perfect intercoder reliability was due to the previous training and that the coders had exact word lists to refer to, when in doubt of a decision. The human-coded results were then compared with computer-coded results. Three iterations of review and word-list refinement were performed. The final average human-to-computer coding agreement was 97%, ranging from 92% to 100% for the 16 issues. 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"). See the appendix for a list of 16 issues and the associated keywords.

In the final analysis, 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.

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 (Tang, Meng, Nguyen, Mei, & Zhang, 2014). 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, whereas 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.[4] If the number of tweets produced by a user is equal to $4q + r$ and r is between

1 and 3, that is, the total number of tweets is not divisible by four, then the last group of r tweets are combined with $4 - r$ preceding tweet(s) to make a set of four. For example, a user who posted seven tweets during the election (i.e., $q = 1$, $r = 3$) would generate two documents with the first four tweets as the first document and the last four tweets as the second document. When, r is not zero, the last two documents of a user may partially overlap. This is, however, limited to an overlap of at most one document per user. Such overlapping documents constitute less than 13% of all documents in our datasets. With this tweet-combining methodology, the Obama-only tweet dataset has a total of 8,069,127 documents and the Romney-only dataset a total of 5,012,680 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.[5] We dropped these non-standard words and characters to conform to the common practices that are followed in previous topic model analyses of Twitter data (e.g., Hong & Davison, 2010; Lim & Buntine, 2014; Mehrotra, Sanner, Buntine, & Xie, 2013; Zhao et al., 2011). But it should be noted that analyzing such information, especially emojis, would be a fruitful direction for future work. 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 with those generated by the dictionary-based analysis. We adopted other parameters as suggested by Řehůřek and Sojka (2010). The parameter initialization recommended in their work is adapted from Hoffman, Bach, and Blei (2010). Specifically, topic weights are symmetrically initialized to the same value for all topics. It would be worthwhile to explore the use of asymmetric topic weights learned from data in future research.

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. Notably, we accounted for the effect of any overlapping tweets in the last two documents of each user. If the last document of a user has $(4 - r)$ tweet(s) in common with the last-but-one document of the user, then its "theta" contribution to the overall topic proportion in the dataset was down-weighted by the fraction $r / 4$.[6]

Comparison and Human Evaluation

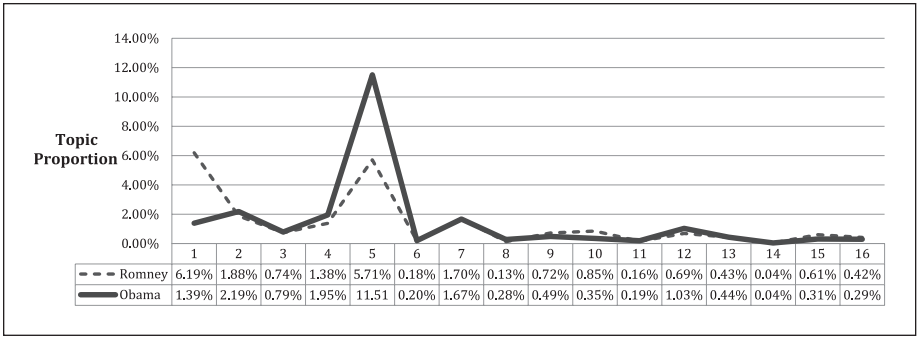
The results generated by the two computer-assisted text analysis methods were compared with each other. To explore the external validity of each method, the two sets of computer-generated results were then compared with human evaluations. A sample of 100 documents was pulled from Obama-only and Romney-only datasets. This sample was not representative of the entire population. It was not our intention to have human coders to evaluate a representative sample of documents, which is logistically impossible. Rather, the primary objective of this step was to use 100 example documents to provide some *qualitative* insight into the possible differences between the two computer-assisted methods. Specifically, in each set of 100 documents, about one third of documents (n = 36) were randomly selected from the entire dataset. Then, 32 documents were randomly selected to represent keyword-based results. That is, 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%). Including documents that represented each computer-assisted method ensured a diverse range of examples to discuss the strength and weakness of the two methods.

Two new independent researchers who had not previously seen the data read the documents and decided the main theme of each document separately. The two coders then discussed their interpretations and reread the documents until they reached a consensus. The human evaluation results were then used to compare with the results generated by the two computer-assisted methods. For each document, the coders discussed and decided which method generated "topic(s)" that were closer to their own evaluation results. For the results of LDA analysis, the coders were instructed to read both labels and the lists of relevant words.

Results

RQ1 asked about the topic proportion in Twitter's coverage of Obama and Romney 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 30,061,046 tweets mentioning "Obama," 80.1% of them did not specify any predetermined topic. Likewise, out of 18,677,277 tweets about "Romney," 80.9% of them contained no topic.

For tweets that did discuss at least one of the 16 predefined 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" (11.51%) was the most salient topic in Twitter's discussion about Obama, followed by "jobs/unemployment" (2.19%) and "economy in general" (1.95%). When it came to the conversation about Romney on Twitter, "tax" (6.19%) and "foreign affairs" (5.71%) were the two most discussed issues. On the contrary, certain topics such as "role of government" were only represented in a very small proportion of tweets about either political candidate.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Romney	6.19%	1.88%	0.74%	1.38%	5.71%	0.18%	1.70%	0.13%	0.72%	0.85%	0.16%	0.69%	0.43%	0.04%	0.61%	0.42%
Obama	1.39%	2.19%	0.79%	1.95%	11.51	0.20%	1.67%	0.28%	0.49%	0.35%	0.19%	1.03%	0.44%	0.04%	0.31%	0.29%

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

Note. The sample includes a total of 30,061,046 tweets about Obama and 18,677,277 tweets about Romney. 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) lesbian, gay, bisexual, transgender(LGBT)/same-sex marriage; (10) abortion; (11) environment/climate; (12) energy; (13) education; (14) role of government; (15) middle class; (16) welfare.

In answering RQ2, the LDA-based analysis automatically discovered 16 most important "topics" inherent in Twitter's coverage for Obama and Romney. 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.

In the case of tweets mentioning Obama, the researchers were not able to determine a coherent, sensible issue for topic 10 (see Table 1). This represented a substantial proportion of tweets (50.18%). Notably, the most frequent words in this topic are mostly common functional words that can be found in any conversation. This can be therefore interpreted as a "background discussion topic." For the rest of the tweets, the most salient topic (15.31%) was about "Foreign affairs," specifically the Benghazi attack. The other 14 LDA-generated topics appeared to be somewhat evenly distributed among the tweets. It is worthwhile to note that one of the most important topics about Obama on Twitter was in Spanish (3.28%), which was twice as much as that about Romney (1.55%).

Table 2 illustrates the LDA-generated topics regarding Romney. Like Twitter's coverage of Obama, about half of the tweets that mentioned "Romney" did not correspond to any meaningful topic. The "presidential debate & fact checking" (11.54%) was the most salient topic on Twitter about Romney. Notably, "uncivil discourse" (5.35%) was the second most prominent "topic," a pattern not found in tweets about Obama. The two tax-related issues—Romney's comment that 47% of Americans pay no income tax (4.25%) and the impact of tax cuts on middle class (4.20%)—were also salient in tweets about Romney.

Table 1. Top 16 "Topics" on Twitter's Coverage of Obama.

"Topic" and associated words	Label	Proportion (n = 30,061,046)
1. term, second, ralli, campaign, unit, photo, nd, state, kati, fiscal, crowd, join, cliff, ap, visit, springsteen, veteran, whitehous, staff, honor, address, pm, volunt, iowa, day, fire, agenda, denver	Obama's reelection campaign & fiscal cliff	2.29%
2. tax, cut, pay, busi, health, women, care, obamacar, student, gas, rais, colleg, million, plan, energi, abort, money, green, small, spend, billion, oil, cost, fund, medicar, price, compani, pac, immigr, teacher, alien, afford	Mixed topics (e.g., tax, health care, education, women, abortion, immigration, energy)	3.49%
3. de, la, que, el, en, un, di, president, se, para, por, los, las, es, le, yo, ya, con, su, il, ha, est, da, fuck, usa, dan, del, al, si, lo, una	Spanish	3.28%
4. debat, middl, presidenti, class, trump, jay, realdonaldtrump, offer, justin, donald, perform, million, east, beyonc, birth, tonight, eat, kenya, record, transcript, colleg, dinner, dollar, moder, certif	Presidential debate & Obama's birth certificate controversy & celebrity endorsement	2.29%
5. job, offic, unemploy, creat, took, number, rate, dennisdmz, approv, month, street, million, sinc, seanhann, wall, pr, candi, golf, halloween, privat, intellig, redistribut, kid, hoe, treat	Jobs/unemployment	2.23%
6. poll, ohio, lead, voter, florida, earli, state, show, gop, among, elector, swing, virginia, va, women, fl, edshow, democrat, predict, republican, pa, colorado, point, counti, wisconsin, key, wi, fivethirtyeight, oh, percent	Election news/horse race	2.70%
7. tcot, gop, teaparti, tlot, nigga, lnyhbt, slone, food, stamp, patdollard, nobama, ocra, sgp	Partisanship & food stamp	1.74%
8. year, bush, economi, blame, four, gop, debt, econom, yrs, ago, fix, trillion, mess, auto, deficit, georg, polici, budget, industri, fail, reagan, rep	Bush year legacy & federal debt	2.98%

(continued)

Table 1. (continued)

"Topic" and associated words	Label	Proportion (n = 30,061,046)
9. youtub, seal, kill, bin, laden, home, navi, osama, movi, war, got, troop, us, iraq, rape, film, end, afghanistan, gave, back, want, credit, better, els, healthcar, palin, facebook	Foreign affairs—Bin Laden	1.66%
10. vote, will, elect, can, get, go, like, say, just, win, peopl, now, don	NA	50.18%
11. news, white, hous, cnn, fox, msnbc, abc, gay, men, nbc, cbs, latino, marriag, burn, break, bodi	Latino voters & LGBT	2.02%
12. campaign, anti, video, ad, big, new, donat, surpris, tv, bird, social, air, use, camp, octob, star, slogan, hollywood, donor	Campaign fundraising & Romney's "big bird" comment	2.94%
13. michell, speech, clinton, bill, dnc, ladi, convent, first, democrat, quot, chair, eastwood, clint, empti, rnc	The Democratic National Convention	2.73%
14. benghazi, lie, american, libya, attack, us, call, administr, media	Foreign affairs–Benghazi	15.31%
15. sandi, hurrican, christi, storm, gov, endors, powel, new, chris, tour, fema, ny, colin, disast, nj, respond, declar, jersey, victim, prais, rush, damag, katrina, cancel, op, cuz, photo, hit, respons, cont, limbaugh, day, quick	Hurricane Sandy	1.77%
16. new, endors, post, victori, elect, washington, reuter, york, blog, layoff, time, bloomberg, employe, ceo, boo, worker, market, tea, reelect, parti, chang, celebr, huffingtonpost, stock, mayor, chavez, climat, lay, nytim, texa, michael, madonna	Endorsement of Obama	2.39%

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. LDA = Latent Dirichlet Allocation; LGBT = lesbian, gay, bisexual, transgender.

Table 2. Top 16 "Topics" on Twitter's Coverage of Romney.

"Topic" and associated words	Label	Proportion (n = 18,677,277)
1. state, sandi, fema, relief, hurrican, disast, donat, moment, storm, use, america, unit, slogan, campaign, teapartycat, american, event, feder, victim, red, keep, privat, kkk, cross, awkward, effort, realiz	Hurricane Sandy	1.93%
2. tax, pay, return, million, cut, paid, releas, hide, rais, rate, year, reid, plan, class, middl, harri, incom, mormon, dollar, trillion, lower	Tax & middle class	4.20%
3. video, new, ad, campaign, comment, post, youtub, percent, daili, washington, blog, mourdock, remark, polit, york, fundrais, dailyko, secret, time, motherjon, tape	Romney's 47% comment	4.25%
4. tcot, gop, endors, romneyryan, parti, teaparti, georg, tlot, republican, regist, christian, lil, tea, paulryanvp, nicki, conserv, des, minaj, leader, honor, ron	Endorsement of Romney & VP nomination	2.49%
5. women, full, gay, binder, woman, pro, right, took, small, children, anti, men, equal, immigr, dinner, marriag, gun, abort, pay, life, decis, child, femal, busi, wing, sex, owner, coupl	Romney's "binders full of women" comment & mixed topics (e.g., LGBT, immigration, abortion, public order)	2.02%
6. debat, plan, lie, big, fact, campaign, last, attack, media, check, night, call	Presidential debate & fact checking	11.54%
7. polici, de, foreign, la, el, en, un, que, candi, al, style, los, usa, crowley, da, test, con, se, experi	Spanish	1.55%
8. fuck, win, becom, bitch, make, ass, gone, colleg, bird, go, even, can, shit, around, get, said, school, parent, chrisrockoz, street, teacher, poof, away, stamp, food, money, lol, kid, field, gonna, danc, im, take, struggl	Incivility	5.35%

(continued)

Table 2. (continued)

"Topic" and associated words	Label	Proportion (n = 18,677,277)
9. middl, machin, militari, tagg, class, war, vote, iran, fraud, secret, israel, ohio, death, son, famili, servic, own, palin, mormon, threat, tie, invest, investig, serv, credit, link, zero, compani, cancel, east, peac, vietnam, card, voter, sarah	Foreign affairs & Romney's ties to voting machine company	1.78%
10. news, fox, game, abc, leader, star, cbs, endors, trend, break, dash, volunt, gon, friday, song, hot, bout, light, stacey, that, yahoo, hunger, nbc, foxnew	NA	1.47%
11. ann, speech, presidenti, pick, gop, republican, rnc, vp, christi, convent, run, chris, candid, twerk, eastwood, announc, offici, clint, campaign, mate, rep, accept, chair, nomin	The Republican National Convention	3.77%
12. job, bain, china, sensata, edshow, creat, auto, jeep, worker, compani, profit, thedailyedg, capit, million, outsourc, bailout, chrysler, ship, employe, ceo, ad, detroit, lie, invest	Bain outsourcing jobs to China & auto bailout	2.88%
13. white, black, hous, joke, rape, racist, control, shirt, birth, akin, spent, ha, hispan, race, abort, latino, todd, commerci, wear, certif, born, tryna, birther	Obama's birth certificate controversy & Latino voters & abortion	1.64%
14. poll, ohio, ralli, lead, state, voter, win, crowd, florida, endors, campaign, pa, show, swing, victori, new, predict, pennsylvania, iowa, elector	Election news/horse race	4.56%
15. governor, bill, breitbartnew, massachusett, john, clinton, sign, dnc, twitchyteam, mccain, sticker, union, michellemalkin, bus, veteran, mass, awesom, king, coal, warn, proud, yard, american, best, gov, threw, protest, communiti, african, troop, kerri, attend, former, wasn	NA	1.79%
16. vote, will, like, say, can, elect, just, get, go, win, peopl, want, don, know	NA	48.79%

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. LDA = Latent Dirichlet Allocation; VP = Vice President; LGBT = lesbian, gay, bisexual, transgender.

It is noteworthy here that our above observations about the "background discussion topic" and about uncivil discourse coalescing into a tight topic are consistent with findings in prior topic-modeling work (e.g., Zhao et al., 2011).

RQ3 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 contrary, 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 predetermined topics with a very limited number of keywords (see the appendix). If a tweet did not contain a keyword, it was dismissed. However, without a predetermined "codebook," the LDA-based analysis discovered a wider variety of topics. Some of these topics were not identified by the preexisting research on elections or the researchers earlier in the project. For example, Obama's birth certificate controversy went "viral" on Twitter. This topic was captured in the LDA analysis for both candidates (i.e., topic 4 for Obama, topic 13 for Romney), but not by the dictionary-based method. Although both methods found "foreign affairs" to be a salient topic in Twitter's coverage of Obama, the LDA analysis revealed more detail. According to Table 1, 15.31% of the tweets were about the Benghazi attack in Libya (topic 14) and 1.66% referred to the killing of Osama Bin Laden (topic 9).

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 nearly 0.20% of the tweets mentioning either candidate discussed the environmental issue, a subject that was not captured by the LDA-based approach. However, it should be noted that, in this study, LDA 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 the environment as well.

What also differentiates the two methods is that, although 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 2 referred to a mixture of topics, and topic 8 contained two issues: Bush year legacy and federal debt. These LDA-generated "topics" with mixed information provide insights into how Twitter users associated different subjects in talking about the given candidate.

RQ4 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 22 documents about Obama and 26 about Romney, neither method captured what the text actually meant.

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	18	25
Neither method captured the main idea of the document	22	26
Dictionary-based analysis captured the main idea of the document but LDA-based analysis did not.	15	10
LDA-based analysis captured the main idea of the document but dictionary-based analysis did not.	41	36
NA (i.e., human coders cannot decipher the content)	4	3

Note. LDA = Latent Dirichlet Allocation.

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 96 documents mentioning Obama that were decipherable by human coders, the LDA-based analysis succeeded in capturing the main idea of nearly two thirds of them (n = 59), whereas the dictionary-based approach captured slightly over a third (n = 33). Likewise, for the 97 readable documents mentioning Romney, the LDA-based analysis aligned with the human coding in 61 documents, whereas the dictionary-based approach made correction decisions for only 35 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:

> mitt romney is now part owner of the company that owns the voting machines to be used in battleground states. #romneytreason mitt romney and his son tagg own the company supplying voting machines to ohio. http://t.co/9xki9x81 tagg romney's connection to drug cartel money laundering.

The LDA analysis determined that a salient topic inherent in this document was about Romney's ties to a voting machine company (topic 9). 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. However, the LDA-based analysis improperly indicated that "#14 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, the dictionary-based analysis determined the document was about "abortion." Consider the following document as another example:

> remember when obama said, "there is no absolute truth—and that's the absolute truth." http://t.co/89dtpbjp rt @danegerus: the obama you don't know http://t.co/shhy0q7s obama's arab-american network rt @silverjingles: muslims have been smiling at their enemies while knifing them in the back since time began. obama set up the killings . . . a variety of #obamabrand items in his cult store. flags and swag: https://t.co/aosyqhm4

In this document, the Twitter user tried to promote the conspiracy theory that Obama is a Muslim. However, both the dictionary-based and LDA-based analysis determined the main topic of the document was "foreign affairs." Both decisions produced misleading information for the final analysis of topic proportion.

Discussion

This article presents an empirical study that investigated and compared two computer-assisted text analysis methods: (a) the dictionary-based analysis, perhaps the most popular automated analysis approaches in social science research; and (b) 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 with 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. In general, 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 also, in fact, more cost-effective than other computer-assisted methods. However, a qualitative assessment of the LDA topics also shows several cases where multiple issues and attributes of candidates were merged into one topic. Sometimes these mergers appeared to make sense (e.g., Bush year legacy and federal debt), but, at other times, no logical link in the topics existed (e.g., Foreign affairs and Romney's ties to voting machine company). For these reasons, LDA topic results need human intervention to avoid these types of errors.

It should be highlighted that, to the best of our knowledge, this present study is the first attempt to validate the efficacy of the LDA model in the context of journalism and mass communication research. Considering its decent performance, future research should consider using this method to analyze mass communication text, especially to process large-scaled social media data. For example, when communication scholars have a big dataset, but are unsure of the topics or attributes that exist inside of it, our results suggest the LDA-based analysis will be more effective than using the most frequently used words to devise topic lists.

The dictionary approach does however still maintain a few use 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." For instance, when a scholar knows they only want to look at one issue (i.e., same-sex marriage), a short list of keywords that focus on the issue will likely retrieve strong initial results.

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 word lists associated with that issue. However, it could be more advantageous to take those word lists and compare them with 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.

It should also be noted that, in addition to the dictionary-based text analysis and unsupervised machine learning introduced in this study, supervised learning provides another option to interpret big social data (e.g., Scharkow, 2013). This method begins with manual content analysis and uses human-coded data to train the computer model. Beyond this, computer scientists and engineers are working on the development of more advanced machine learning algorithms, hoping to 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). Again, as we argue in this article, the efficacy and validity of using any advanced computer-assisted methods in the context of journalism and mass communication research begs further analysis.

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.

Increasingly, journalism and mass communication scholars are turning to data that have been processed by a computer (e.g., for sentiment, or presence of an issue). When computer programs are used to annotate or code data, the exact methodologies of how the computer reached a judgment on the data is crucial to ensuring results are valid. Communication scholars often cite the use of programs without detail as to how they work. This is likely because those methodological steps are not clear or available to the researcher. When data are annotated or aggregated, clear methodology is needed. "Black box" methodologies exist in communication scholarship, and this article is a stark warning as to why that can be problematic.

Even when scholars write their own code to annotate data, journal articles remain inadequate repositories for computer scripts (e.g., python code). Beyond this, big datasets remain hard to share. Legal issues and proprietary claims can often restrict researchers from posting online. For these reasons, 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.

Appendix

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"]

Acknowledgments

The authors would like to thank Lauren Kiefer, Fengzhou Sun, and Kate Mays for assisting with the data analysis.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) declared receipt of the following financial support for the research, authorship, and/or publication of this article: The authors received support from the US AFOSR under award number #FA9550-10-1-0458 (subaward # A1795) and the US NSF under award numbers #1218992 and #1527618. The views and conclusions contained in this article are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the agencies.

Notes

1. Moving beyond the "bag of words" assumption, researchers (e.g., Titov & McDonald, 2008; Wallach, 2006) have also studied topic models that are based on bigram (i.e., word pairs) and multi-gram (i.e., a sequence of multiple words).
2. The reason we eliminated tweets that contained both "Obama" and "Romney" (n = 11,573, 232; 15%) was because it is difficult for a computer-assisted analysis to distinguish between Obama-related and Romney-related discussion in tweets that included both terms. Take the example of the following tweet. "Romney acknowledged the anniversary of 9/11 . . . Guess what Obama acknowledged? Oh election is coming up . . . " Although a human coder may find it straightforward to determine that Romney was associated with the issue of 9/11 and Obama was associated with the election, it requires a sophisticated algorithm to automate the decision. This is, however, beyond the scope of this analysis and it is a limitation of the study. However, it should also be noted that the main purpose of this study was to compare the efficacy and validity of the two computer-assisted methods rather than to exhaustively examine how people discussed the two candidates on Twitter. Holding the data constant for both computer-assisted analyses, the study remains valid.
3. In each group, users who produced less than four tweets were not included in the analysis. A total of 9,599,373 (24%) Obama-only tweets and 5,567,517 (23%) Romney-only tweets were eliminated in this step.
4. In the pretesting stage, we produced LDA topic matrices by combining different numbers of consecutive tweets (n =1-10) authored by the same user into a document. We observed that four tweets per document produced the most coherent topic matrices. That is, "topics" were distinct from each other and semantically meaningful. When n is too small, documents do not contain enough information to reliably estimate topics. When n is too large, however, tweets from longer timespans and that contain many different subjects will be combined into one

document. As a result, the generated "topics" will, to a great extent, overlap with each other. Future research should consider alternative ways to combine tweets, for example, combining tweets that follow the same hashtag (Lim & Buntine, 2014; Mehrotra et al., 2013), or tweets from similar users from a similar timeframe (Bak, Lin, & Oh, 2014).

5. We removed the standard set of English stop words (e.g., a, for, the) provided by the "tm" package for R and Twitter-related stop words found in our sample: description, null, text, url, text, href, rel, nofollow, false, true, rt.
6. This method to estimate topic proportions is only approximate. There exist more principled and accurate methods, which can estimate topic tokens for each word or a group of words. These methods are supported in Gensim, MALLET, and other toolkits for topic modeling. Our method was chosen for its simplicity and computational efficiency.

Reference

Bak, J., Lin, C. Y., & Oh, A. (2014, October 25-29). *Self-disclosure topic model for classifying and analyzing Twitter conversations*. In B. Pang & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1986-1996). Doha, Qatar: Association for Computational Linguistics.

Beyer, M. A., & Laney, D. (2012). *The importance of "big data": A definition*. Stamford, CT: Gartner.

Blei, D. M., & Lafferty, J. (2006). Correlated topic models. *Advances in Neural Information Processing Systems*, *18*, 147-154.

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

Chuang, J., Roberts, M. E., Stewart, B. M., Weiss, R., Tingley, D., Grimmer, J., & Heer, J. (2015). TopicCheck: Interactive alignment for assessing topic model stability. In Y. Weiss, B. Schölkopf, & J. C. Platt (Eds.), *Proceedings of NAACL-HLT* (pp. 175-184). Cambridge, MA: MIT Press.

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*, 862-883.

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

Goel, S., Anderson, A., Hofman, J., & Watts, D. (2013). The structural virality of online diffusion. *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*, 811-824.

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for Latent Dirichlet Allocation. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, & A. Culotta (Eds.), *Advances in neural information processing systems* (pp. 856-864). Red Hook, NY: Curran Associates.

Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in twitter. In P. Melville, J. Leskovec, & F. Provost (Eds.), *Proceedings of the first workshop on social media analytics* (pp. 80-88). New York, NY: ACM.

Hu, Y., Boyd-Graber, J., Satinoff, B., & Smith, A. (2014). Interactive topic modeling. *Machine Learning*, *95*, 423-469.

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

Kiousis, S. (2004). Explicating media salience: A factor analysis of New York Times issue coverage during the 2000 US presidential election. *Journal of Communication*, *54*, 71-87.

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

Laney, D. (2001). *3D data management: Controlling data volume, velocity and variety* (META Group Research Note). Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Leccese, M. (2009). Online information sources of political blogs. *Journalism & Mass Communication Quarterly*, *86*, 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*, 34-52.

Lim, K. W., & Buntine, W. (2014). Twitter opinion topic model: Extracting product opinions from tweets by leveraging hashtags and sentiment lexicon. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (pp. 1319-1328). New York, NY: ACM.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, *28*, 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*, 76-91.

Manning, C., Raghavan, P., & Schütze, H. (2009). Flat clustering. In *Introduction to information retrieval* (pp. 349-374). New York, NY: Cambridge University Press.

McCombs, M. (2014). *Setting the agenda* (2nd ed.). Cambridge, UK: Polity Press.

Mehrotra, R., Sanner, S., Buntine, W., & Xie, L. (2013). Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (pp. 889-892). New York, NY: ACM.

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

Newman, D., Chemudugunta, C., Smyth, P., & Steyvers, M. (2006). Analyzing entities and topics in news articles using statistical topic models. In S. Mehrotra, D. D. Zeng, & H. Chen (Eds.), *Intelligence and security informatics* (pp. 93-104). Berlin, Germany: Springer.

Petrocik, J. R. (1996). Issue ownership in presidential elections, with a 1980 case study. *American Journal of Political Science*, *40*, 825-850.

Rajadesingan, A., Zafarani, R., & Liu, H. (2015). Sarcasm detection on Twitter: A behavioral modeling approach. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (pp. 97-106). New York, NY: ACM.

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modeling with large corpora. In N. Calzolari et al. (Eds.), *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45-50). Valletta: University of Malta.

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

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

Scharkow, M. (2013). Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. *Quality & Quantity*, *47*, 761-773.

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*, 113-116.

Tang, J., Meng, Z., Nguyen, X., Mei, Q., & Zhang, M. (2014). Understanding the limiting factors of topic modeling via posterior contraction analysis. In E. P. Xing & T. Jebara (Eds.), *Proceedings of the 31st International Conference on Machine Learning* (pp. 190-198). New York, NY: ACM.

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

Titov, I., & McDonald, R. (2008). Modeling online reviews with multi-grain topic models. In W.-Y. Ma, A. Tomkins, & X. Zhang (Eds.), *Proceedings of the 17th international conference on World Wide Web* (pp. 111-120). New York, NY: ACM.

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*, 296-316.

Wallach, H. M. (2006). Topic modeling: Beyond bag-of-words. In W. Cohen & A. Moore (Eds.), *Proceedings of the 23rd international conference on Machine learning* (pp. 977-984). New York, NY: ACM.

West, M. D. (Ed.). (2001). *Theory, method, and practice in computer content analysis*. Westport, CT: Ablex.

Xiang, D. (2013). China's image on international English language social media. *Journal of International Communication*, *19*, 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*, 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, Germany: Springer.

Author Biographies

Lei Guo is currently an assistant professor at Boston University. She earned her PhD from the University of Texas at Austin in 2014. Her research focuses on the development of media effects theories, emerging media technologies and democracy, and international communication. Her expertise also includes the use of computational methods such as machine learning and network analysis to analyze big social data.

Chris J. Vargo (PhD, University of North Carolina at Chapel Hill) is an assistant professor of public relations at The University of Alabama. He specializes in the use of computer

science methods to investigate social media using theories from the communication and political science disciplines. His research methods of specialization include text mining, machine learning, computer-assisted content analysis, data forecasting, information retrieval, and network analysis. He has published in the *Journal of Communication* and *Mass Communication & Society*.

Zixuan Pan recently received the MA degree in electrical and computer engineering from Boston University where he also worked as a research fellow. He is currently a machine-learning engineer in Yodlee's research team. His current research interests are in machine learning and data mining.

Weicong Ding received his PhD from the Department of Electrical and Computer Engineering in Boston University in 2015, and his BSc degree in Electrical Engineering from Tsinghua University, Beijing, China, in 2010. He is currently a research scientist in Techinicolor Research. His current research interests are in user analytics, machine learning, and data mining.

Prakash Ishwar received the BTech degree in electrical engineering from the Indian Institute of Technology, Bombay in 1996 and the MS and PhD degrees in electrical and computer engineering (ECE) from the University of Illinois Urbana–Champaign in 1998 and 2002, respectively. After two years as a post-doctoral researcher in the electrical engineering and computer sciences department at the University of California, Berkeley, he joined the faculty of Boston University where he is currently associate professor of ECE. His current research centers on data science to advance statistical and computational tools for learning and inference problems using both model-based and data-driven methods. He is a recipient of the 2005 NSF CAREER award, a co-recipient of the AVSS'10 best paper award, and a co-winner of the 2010 ICPR Aerial View Activity Classification Challenge.