Content Analysis and the Algorithmic Coder: What Computational Social Science Means for Traditional Modes of Media Analysis

By
RODRIGO ZAMITH
and
SETH C. LEWIS

To deal with ever-larger datasets, media scholars are increasingly using computational analytic methods. This article focuses on how the traditional (manual) approach to conducting a content analysis—a primary method in the study of media messages—is being reconfigured, assesses what is gained and lost in turning to computational solutions, and builds on a "hybrid" approach to content analysis. We argue that computational methods are most fruitful when variables are readily identifiable in texts and when source material is easily parsed. Manual methods, though, are most appropriate for complex variables and when source material is not well digitized. These modes can be effectively combined throughout the process of content analysis to facilitate expansive and powerful analyses that are reliable and meaningful.

Keywords: content analysis; computational social science; computational content analysis; hybrid content analysis; digital research methods; media analysis; algorithms

The abundance of digitized data has become a defining feature of modern society, and particularly of communication that is expressed through digital, social, and mobile platforms: tweets, likes, links, shares, texts, posts, tags, and more—literally billions of data points about social behavior that, potentially, might be assembled, accessed, and ultimately analyzed by various institutions and individuals. For

Rodrigo Zamith is an assistant professor in the Journalism Department at the University of Massachusetts, Amherst. His research focuses on the reconfiguration of journalism in a changing media environment as well as the development of digital research methods.

Seth C. Lewis is an assistant professor and the Mitchell V. Charnley Faculty Fellow in the School of Journalism and Mass Communication at the University of Minnesota–Twin Cities. His research focuses on journalism and technology. He is coeditor of Boundaries of Journalism: Professionalism, Practices, and Participation *(Routledge 2015).*

DOI: 10.1177/0002716215570576

communication and media research, especially, the possibilities are great: as computational tools and processes have become easier to employ (e.g., via open-source software), as large-scale datasets of digital media content have become more readily attainable (e.g., by scraping tweets or websites), and even as print media content, such as old newspapers and books, have become increasingly available in digital form, new types of large-scale, algorithm-driven analyses of media content have become possible, enabling scholars to address novel questions. To cite just one example, Colleoni, Rozza, and Arvidsson (2014) used more than a billion data points to reveal key differences in the structures of political homophily among Democrats and Republicans on social networks such as Twitter. Such "naturally occurring" data such as public tweets, and the growth in computing capacity that facilitates the collection and analysis of such voluminous data, marks a key turn toward *computational social science* (Shah, Cappella, and Neuman, this volume).

As a distinct approach to social inquiry, computational social science is characterized by research that (1) uses large, complex datasets; (2) often involves social and digital media sources; (3) employs algorithmic or computational solutions to generate patterns and inferences from data; and (4) is applicable to social theory in a wide variety of domains (Shah, Cappella, and Neuman, this volume). Examples of such research may be found across a range of disciplines, including a growing number and variety at the intersection of the social sciences and the digital humanities (Bruns 2013). Much of this work involves the quantitative analysis of textual content. Yet unlike traditional content analyses—which rely predominantly on human judgments—these studies are largely driven by algorithms and frameworks that seek to automate the coding process. With this in mind, we ask, What does this turn toward computational social science mean for traditional forms of content analysis?

That question speaks to the very future of content analysis as a method—one of the primary methods of mass communication research for many decades, and one considered essential to a scientific study of communication (Riffe, Lacy, and Fico 2014). The purpose of this article, therefore, is to (1) consider the traditional way of conducting content analysis in light of the algorithmic coder, (2) assess what is gained and lost in turning to purely algorithmic solutions, and (3) discuss an alternative approach that leverages traditional and computational approaches in tandem. Such an approach, we argue, can facilitate more expansive and powerful—yet still reliable and meaningful—forms of content analysis within the present turn toward computational social science.

Content Analysis and Algorithmic Approaches

Riffe, Lacy, and Fico (2014, 19) present what is, in our opinion, a particularly good definition of *quantitative* content analysis: "the systematic and replicable examination of symbols of communication, which have been assigned numeric values according to valid measurement rules, and the analysis of relationships involving those values using statistical methods, to describe the communication, draw inferences about its meaning, or infer from the communication to its context, both of production and consumption." They also present a comprehensive framework for conducting content analysis—in a traditional fashion—breaking the process up into three segments: (1) conceptualization and purpose, (2) design, and (3) analysis.

We focus here on four key processes within that overall procedure: (1) the development of the coding protocol and sheet; (2) the specification of the population and, if applicable, the sample; (3) the establishment of intercoder reliability; and (4) the coding of content. The other steps in the process, such as identifying the problem and reviewing the relevant literature, we argue, are generic, in the sense that they are applicable to most scholarly work, regardless of the quantitative method employed. We also wish to clarify that the use of algorithms to code content is not a new phenomenon (for an example, see Stone et al. 1962). However, what is novel about current efforts is the desire to automate virtually the entire procedure—that is, to perform a content analysis with minimal human intervention. We thus proceed to compare traditional and computational forms of content analysis, explicating the human-centric and machine-centric steps associated with each to set up a broader discussion about the relative benefits and drawbacks of the algorithmic coder, the computational counterpart to a human coder.

Coding protocol and the code sheet

Traditionally, once the hypotheses and research design have been finalized, the researcher must develop a coding protocol, or a set of explicit and detailed coding instructions for a human coder to follow. This, in turn, requires the researcher to balance the level of instruction to ensure that it is neither so detailed that it makes the application of the instruction too narrow and the protocol too complex, nor so vague that it leaves too much room for interpretation and thereby undermines reliability. A simple code sheet should then be developed for the coder, listing every variable being coded and leaving some room for the coder to enter his or her code.

With an algorithmic coder, there is no room for ambiguity in the coding protocol. For example, a dictionary-based approach necessitates the development of extensive lists of all possible permutations of distinct units—typically words and their combinations—that represent a given construct (e.g., a victimization story frame; see Vliegenthart and Roggeband 2007). Machine learning approaches require the development of clearly specified models to identify patterns and make inferences about the object of study from the data, typically by calculating the probability that content is of a certain class and categorizing it based on the highest probability (e.g., Grimmer 2010). The algorithmic coder then employs these lists of words and models with the instructions for how to use them coming in the form of unambiguous computer code.

The lack of ambiguity in an algorithmic approach sets it apart from traditional content analysis for two reasons. First, it requires the researcher to be acutely aware of the many ways that a construct may manifest itself in a given text, especially when utilizing a dictionary-based approach. While researchers have long needed such awareness to develop comprehensive codebooks in traditional content analysis, they could rely on the judgment of human coders to adapt to unusual manifestations; under an algorithmic approach, this is not possible. Second, it may potentially facilitate replicability and transparency by ensuring that every step along the way is clearly documented, and further enables content analyses to be more comparable by virtue of the adoption of the same dictionaries and models. This is a stark contrast to traditional content analysis, where a considerable amount of instruction occurs verbally during the coder training process, with only a portion of that instruction making it into the protocol.

Additionally, an algorithmic approach departs from traditional analysis by being considerably more iterative. Algorithms and dictionaries must often be repeatedly revised and tweaked to improve their performance. While it is not uncommon for content analysts to produce a handful of revisions to a coding protocol, an algorithmic approach may involve dozens of rounds of changes to ensure that the classification of items yields a satisfactory level of construct validity.

Finally, the algorithmic coder does not need a code sheet. As content is coded, data are automatically added to digital, structured datasets that can be easily imported into a statistical analysis program. This point, while perhaps obvious to some, is worth noting because of its significance for validity and reliability. The algorithmic process removes data-entry error, both on the part of a coder mislabeling text and as data are transferred from a code sheet to the final dataset.

Specifying the population

The specification of the population requires the researcher to decide which materials to analyze and how much of those materials to analyze. Here, the population comprises all the potential content that falls within the restrictions set by the researcher during the design process. Researchers may, and often do, opt to use a sample, or portion, of that population to reduce the workload of the coders, or to enable them to loosen other restrictions (e.g., look at more media outlets or over a longer period of time).

An algorithmic approach departs from traditional content analysis in that it can generally be scaled up with ease. Provided that vocabularies, features, and patterns do not change significantly (e.g., selecting texts that vary greatly in structure and conventions), the difference between coding a large corpus of material and a very large corpus of material is relatively minor (Hopkins and King 2010). Researchers may thus use a larger sample, which is generally more likely to represent the overall population, or even a census.[1] Additionally, an algorithmic approach may leverage computer scripts to systematically locate, obtain, and organize data, ensuring that the population—or an appropriate sample—is captured more systematically than if done by a human being (Karlsson and Strömbäck 2010). Last, there is often a need to preprocess collected content and convert it into research-grade data that are in a form that may be easily parsed and analyzed by an algorithm. For example, this may include identifying and correcting repeated errors in source documents, such as words that were split incorrectly during optical character recognition (see Leetaru 2012)—a painstaking process that has only limited parallels to a traditional approach.

Assessing intercoder reliability

Once the coding protocol has been designed and the population specified, the researcher will want to ensure that the coding protocol's definitions are reliable. Typically, this is done through the proxy of having multiple individuals double-code a randomly selected subset of the sample. The researcher then assesses whether the coders made the same coding decisions the majority of the time, often using statistics such as Scott's pi or Cohen's Kappa. Generally, multiple rounds of reliability testing are necessary to attain acceptable coefficients.

Because computers are deterministic machines, they are able to execute a given set of instructions with perfect reliability (Grimmer 2010). There is no need to assess intercoder reliability in an algorithmic approach, thus distinguishing it from traditional content analysis. This is noteworthy in light of scholars' concern about the poor reporting of intercoder reliability in published content analyses (Lovejoy et al. 2014; Riffe and Freitag 1997).

However, to assess the *validity* of an algorithm, researchers often choose to measure it against a "gold standard"—typically a human-coded dataset that is presumed to represent the "correct" coding decisions (Grimmer and Stewart 2013). Additionally, the chosen algorithm may also be compared against competing ones in a benchmarking process (for an example, see Thelwall et al. 2010). There is no counterpart to this procedure in traditional content analysis: the decisions made by human coders operating under one codebook are rarely compared against those made under comparable codebooks, and the validity of a codebook is seldom established by measuring it against a specific standard.

Coding materials

Once reliability has been established, the remainder of the sample or census is coded. Sometimes, multiple coders—all of whom should have participated in the assessment of reliability—will divvy up the work; otherwise, a single coder will take on the remaining work. At this point, no major changes should be made to the coding protocol since it would necessitate a reassessment of its reliability. Data are then stored, analyzed, and interpreted.

An algorithmic approach departs from traditional content analysis here by being exponentially faster. Simple analyses that may take human coders months can often be accomplished in minutes or hours using a desktop computer (Manovich 2012). Additionally, more sophisticated analyses can often be parallelized. Rather than adding human coders to expedite the analysis, researchers may quickly add dozens more central processing units (CPUs) at any point along the way, provided they have the resources, to speed up the process. Consequently, recoding items—perhaps in response to suggestions by peers or reviewers—is generally both easy and expeditious with the algorithmic coder. This is a stark contrast to traditional content analysis, wherein such modifications are generally infeasible.

An algorithmic approach to content analysis, therefore, does not neatly fit into the traditional framework for conducting a content analysis. Unlike traditional content analysis, an algorithmic approach requires the coding protocol and instructions be unambiguous, be more iterative, be conducted on a far grander scale, and be considerably more flexible in accommodating post-hoc adjustments. Moreover, the absence of the need for a code sheet or to perform intercoder reliability, the need to preprocess content to ensure that it is in a format that is amenable to machine processing, and the common requirement of "validating" algorithms by measuring them against a gold standard all make it difficult to map the work of the algorithmic coder onto the traditional framework for content analysis. A revised framework for the content analyst operating in a computational social science environment is therefore warranted (see also Herring 2010).

Challenges for Algorithmic Content Analysis

Communication researchers, especially those interested in studying human perceptions and practices as expressed on digital and social media platforms, currently have access to a growing variety and volume of data. This "siren-song of abundant data," as Karpf (2012, 648) has called it, is enticing on one hand, but vexing on the other. Much of the public data available to researchers are ephemeral and thus hard to capture reliably; they are often polluted by "noise" from the influence of spammers (e.g., appropriating a popular hashtag with unrelated information); and they are often more limited than they may appear (e.g., Twitter allowing access to only a portion, and not all, of its publicly available tweets via its application program interface [API]). All of this leads Karpf to encourage "methodological skepticism" because "the glittering promise of online data abundance too often proves to be fool's gold" (2012, 652; see related discussion in boyd and Crawford 2012).

What do such cautions mean for content analysis in particular? In light of the considerable benefits offered by algorithmic approaches—namely, the potential to quickly collect and analyze massive amounts of digital content with perfect reliability and exceptional transparency—many researchers may be tempted to eschew traditional content analysis for an algorithmic approach in a computational social science environment. However, in comparison to the algorithmic coder, a human coder in many cases may be more attuned not only to the richness and context of the topic at hand, but also to the data-quality issues described by Karpf (2012). A human, for instance, may be able to better recognize the presence and relative pervasiveness of spam when studying tweets that include a particular hashtag. Ultimately, humans can understand the larger sociocultural contexts through which tweets, like other social media posts, function as a form of communication (see discussion in Weller et al. 2014).

Beyond issues of data quality, as Krippendorff (2013, 210) notes, “programming a machine to mimic how humans so effortlessly understand, interpret, and rearticulate text turns out to be an extraordinarily difficult, often impossible, undertaking.” While this may overstate the near-term aims of researchers in this area, it points to the broader ambition of limiting the role of human coders. Specifically, Krippendorff points to the difficulty of attaining acceptable levels of validity with complex and oftentimes ambiguous textual representations, such as sarcastic remarks and metaphors.[2] As several scholars have argued, while algorithmic approaches yield satisfactory results in surface-level analyses or analyses that focus on structural features, their performance is significantly worse when assessing more complex features of texts (Conway 2006; Sjøvaag and Stavelin 2012).

A Hybrid Approach for Computational Social Science

Recognizing the current challenge of using algorithms to accurately analyze and classify complex human communication, Lewis, Zamith, and Hermida (2013, 36) proposed a hybrid approach to content analysis “that combines computational and manual methods throughout the process . . . [to] retain the strengths of traditional content analysis while maximizing the accuracy, efficiency, and large-scale capacity of algorithms for examining Big Data.” In their analysis of news sourcing practices on Twitter during the Arab Spring, they used computational methods first to organize the data via a Python script, which turned a messy text file containing tens of thousands of tweets into a standardized, comma separated values (CSV) data file through which key variables could be identified and studied for initial patterns. Then, additional computer scripts allowed for the coding of simple features such as the usernames mentioned in a particular tweet—a necessary variable for understanding how certain individuals were included in sourcing the news (Hermida, Lewis, and Zamith 2014). Thereafter, Lewis and colleagues developed an electronic interface to facilitate the work of human coders and thereby reduce—or even eliminate—certain sources of error. “Through it all,” they emphasize, “computational means were enlisted to enhance, rather than supplant, the work of human coders, enabling them to tackle a larger body of data while remaining sensitive to contextual nuance” (Lewis, Zamith, and Hermida 2013, 47).

In a similar vein, Sjøvaag and colleagues (Sjøvaag, Moe, and Stavelin 2012; Sjøvaag and Stavelin 2012) utilized computer scripts to “freeze the flow of online news” (Karlsson and Strömbäck 2010, 16) to facilitate a hybrid form of content analysis. They first used Python scripts to scrape a year’s worth of coverage—nearly seventy-five thousand news articles produced by the Norwegian Broadcasting Company (NRK)—and automatically code for web-specific features such as hyperlinks and multimedia. Thereafter, in a more traditional fashion, they manually coded a smaller subset of articles to capture contextual features such as topics, themes, and frames—none of which could be adequately classified via an algorithm (for details, see Sjøvaag and Stavelin 2012).

The development of computational tools and frameworks that can facilitate the blending of human judgment and algorithmic efficiency strikes us as an area of research that deserves additional attention from content analysts amid the turn toward computational social science. This suggestion may appear at first to be paradoxical: after all, how can researchers analyze massive, complex sets of textual data with human involvement? We concur that there are instances where human involvement is simply infeasible. However, we also argue that in a considerable amount of scholarly applications, human involvement is entirely practical. First and foremost, we share the position that a census is not always necessary. Just because researchers can conduct a census does not mean that they should eschew the tradition of sampling that has served the field well (Riffe, Lacy, and Fico 2014). As Mahrt and Scharkow (2013, 20) put it, "researchers need to consider whether the analysis of huge quantities of data is theoretically justified, given that it may be limited in validity and scope, and that small-scale analyses of communication content or user behavior can provide equally meaningful inferences. . . ."

However, in some research, even a good sampling strategy will yield a sizable corpus of textual data. Consider a hypothetical study of the sourcing practices of journalists at dozens of news outlets in stories about immigration over a 30-year period. Such a study may involve the analysis of hundreds of thousands of full-length stories—far more than a small group of human coders could analyze in a reasonable amount of time. While an algorithmic approach may appear to be ideal for such a research endeavor, it is probable that certain variables, such as source type, would yield results that are lacking validity because of the limited amount of explicit attribution information in news articles.

A hybrid approach, however, would make great sense for such a project. Algorithms could be leveraged to code for simple variables such as the outlet from which the article was found and the date of the article. In such instances, Named Entity Recognition[3] (see Béchet 2011) could be leveraged to identify sources within those articles, and dictionary-based or machine learning–based approaches could be used to estimate the appropriate classification of complex variables. Additionally, an electronic interface could then be leveraged to allow coders to quickly verify algorithmic decisions by prefilling code sheets, with visual cues added to the source document to illustrate the rationale for the algorithm's decision.

To further illustrate this point using the aforementioned source-type variable, consider the possibility of using algorithms to capture every text segment in which a specific source appears, presenting these segments sequentially to a human coder and conveying the estimated probability for each source-type category, and automatically propagating the human's single coding decision for every article in which that source appears. While some sources may only appear once in the entire corpus of coverage (e.g., a man-on-the-street source), a considerable amount are likely to reappear multiple times (e.g., government officials).

While a myriad of additional examples could be readily offered, this simple scenario helps to illustrate how a blend of computational tools and human expertise could reduce the likelihood of having invalid coding decisions—a common criticism of the algorithmic approach (Manovich 2012)—while enabling researchers to tackle far larger datasets than they otherwise could. This, we argue, enables content analysis to remain relevant and yield insightful knowledge within a computational social science paradigm. However, while the turn toward computational social science has excited a flurry of activity in the development of more sophisticated techniques for modeling language and classifying phenomena (e.g., Grimmer 2010), considerably less work has been done in the development of tools and frameworks that facilitate the interaction of computational tools and human expertise.[4]

Conclusion

The turn toward computational social science requires scholars to reconsider how content analysis is done and whether it needs to be adapted to remain relevant in a changing research environment. To deal with the large, complex datasets that characterize this turn, a number of scholars have turned to computational forms of content analysis and subsequently shifted from the human coder to the algorithmic coder (e.g., Grimmer 2010; Vliegenthart and Roggeband 2007). The work of the algorithmic coder and the design of an algorithmic approach do not neatly fit into the traditional framework for conducting content analysis. While some processes, such as the development of a coding protocol, require only reconceptualization, other processes, such as the need to validate algorithmic decisions against a gold standard, hardly correspond at all with that traditional framework. While we would not go so far as to argue that such forms are entirely distinct, we do believe that a revised framework for the content analyst operating in this new environment is needed (cf. Herring 2010).

To be sure, computational approaches and the algorithmic coder may yield great benefits, such as increased efficiency, transparency, and post-hoc malleability. Such benefits cannot be ignored. Nevertheless, computational approaches also have clear limitations. Their reliance on digital materials that are often of questionable quality and their focus on relatively simple and unambiguous content—lest they run the risk of producing results lacking in validity—may be problematic (Conway 2006; Mahrt and Scharkow 2013). As such, we argue that for content analysis to remain relevant in the turn toward computational social science, a hybrid approach must be further developed, one that preserves the contextual sensitivity and validity that are central to traditional content analysis and combines it with the large-scale capacity and reliability of computational approaches. Specifically, while the work of Lewis and colleagues (2013) and Sjøvaag and colleagues (Sjøvaag, Moe, and Stavelin 2012; Sjøvaag and Stavelin 2012), among others, offer a good starting point, we believe that there are several avenues for building on their work. To this end, we advocate for the development of tools and frameworks that facilitate the interaction of computational tools and human expertise.

In effect, we are arguing not only for preserving the best of the old human coder while also embracing the best of the new algorithmic coder, in an idealized or abstract sense, but also that it is time to take tool-building more seriously within our discipline. We not only need to forge conceptual ways of thinking about these problems but also build better technical systems to conduct the kind of content analysis that truly blends the best of both worlds, human and machine alike. Such a system might not only do what we are already accustomed to having algorithms do (e.g., automatically coding certain machine-readable structural characteristics) but also leverage advanced computational techniques to provide cues for human coders and facilitate their work, thereby leading to faster, more reliable, and ultimately more valid coding decisions, even while retaining the important contextual awareness that humans bring to the equation. Thus, the human coder and algorithmic coder, working together, may help communication researchers to keep content analysis relevant in the turn toward computational social science and also in fact propel it forward as a method, such that it will, perhaps, be as indispensable to the next generation of communication researchers as it has been to this one.

Notes

[1] As other scholars have noted, turning to larger sample sizes or a census may not necessarily yield better inferences (Mahrt and Scharkow 2013). Put differently, a proper sampling technique may yield inferences that are just as good as if one had looked at the entire sample (Riffe, Lacy, and Fico 2014). Additionally, as a practical concern, it is generally inappropriate to utilize inferential statistics when analyzing a census (and is oftentimes meaningless when looking at massive datasets; see Ruggles 2014), though many reviewers nevertheless expect to see them and sometimes reject manuscripts for failing to use them. However, in some instances it is necessary to analyze a census to obtain a complete picture (e.g., to not miss out on key bridging nodes when generating a network of actors from textual material). When dealing with large sample sizes or a census, it is especially important that researchers focus on the practical significance of differences and effects, and not be overreliant on statistical significance (Berman 2013).

[2] Béchet (2011, 261) offers a simple illustration by discussing the challenge of having a Named Entity Recognition algorithm correctly disambiguate, without prior specification, the English football club Sheffield Wednesday (an organization) from Sheffield (a location) and Wednesday (a day of the week).

[3] Named Entity Recognition refers to “a task consisting of detecting segments of a document . . . expressing a direct reference to a unique identifier” (Béchet 2011, 257). It often uses linguistic models to computationally identify individuals, organizations, and locations within a given text.

[4] There are existing tools such as Crimson Hexagon that facilitate the collection of data and allow researchers unfamiliar with techniques in the fields of natural language processing to make use of advanced machine learning algorithms. While such tools may indeed be of great use to some researchers, they also introduce some potential problems for scientific research. First, most user-friendly tools are restricted to a certain kind of data, typically social media data and often just Twitter. This restricts the questions and phenomena social scientists may study. Second, most of these tools—especially user-friendly, all-in-one solutions—are closed-source and often offer limited capacity for exporting important information necessary to replicate (or extend) a study. This is especially true for web-based services, which may change their proprietary algorithms and offer no recourse for using a previous algorithm. Thus, while such tools can be beneficial for many researchers, their use should be carefully considered.

References

Béchet, Frédéric. 2011. Named entity recognition. In *Spoken language understanding*, eds. G. Gokhan Tur and Renato De Mori, 257–90. West Sussex, UK: John Wiley & Sons, Ltd.

Berman, Jules J. 2013. *Principles of big data: Preparing, sharing, and analyzing complex information*. Waltham, MA: Morgan Kaufman.

boyd, danah, and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15 (5): 662–79.

Bruns, Axel. 2013. Faster than the speed of print: Reconciling "big data" social media analysis and academic scholarship. *First Monday* 18 (10). doi:10.5210/fm.v18i10.4879.

Colleoni, Elanor, Alessandro Rozza, and Adam Arvidsson. 2014. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of Communication* 64 (2): 317–32.

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

Grimmer, Justin. 2010. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis* 18 (1): 1–35.

Grimmer, Justin, and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21 (3): 267–97.

Hermida, Alfred, Seth C. Lewis, and Rodrigo Zamith. 2014. Sourcing the Arab Spring: A case study of Andy Carvin's sources on Twitter during the Tunisian and Egyptian revolutions. *Journal of Computer-Mediated Communication* 19 (3): 479–99.

Herring, Susan C. 2010. Web content analysis: Expanding the paradigm. In *International handbook of internet research*, eds. Jeremy Hunsinger, Lisbeth Klastrup, and Matthew Allen, 233–49. New York, NY: Springer.

Hopkins, Daniel J., and Gary King. 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science* 54 (1): 229–47.

Karlsson, Michael, and Jesper Strömbäck. 2010. Freezing the flow of online news: Exploring approaches to the study of the liquidity of online news. *Journalism Studies* 11 (1): 2–19.

Karpf, David. 2012. Social science research methods in Internet time. *Information, Communication & Society* 15 (5): 639–61.

Krippendorff, Klaus. 2013. *Content analysis: An introduction to its methodology*. Thousand Oaks, CA: Sage Publications.

Leetaru, Kalev H. 2012. *Data mining methods for the content analyst: An introduction to the computational analysis of content*. New York, NY: Routledge.

Lewis, Seth C., Rodrigo Zamith, and Alfred Hermida. 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.

Lovejoy, Jennette, Brendan R. Watson, Stephen Lacy, and Daniel Riffe. 2014. Assessing the reporting of reliability in published content analyses: 1985–2010. *Communication Methods and Measures* 8 (3): 207–21.

Mahrt, Merja, and Michael Scharkow. 2013. The value of big data in digital media research. *Journal of Broadcasting & Electronic Media* 57 (1): 20–33.

Manovich, Lev. 2012. Trending: The promises and the challenges of big social data. In *Debates in the digital humanities*, ed. Matthew K. Gold, 460–75. Minneapolis, MN: University of Minnesota Press.

Riffe, Daniel, and Alan Freitag. 1997. A content analysis of content analyses: Twenty-five years of *Journalism Quarterly*. *Journalism & Mass Communication Quarterly* 74 (3): 515–24.

Riffe, Daniel, Stephen R. Lacy, and Frederick Fico. 2014. *Analyzing media messages: Using quantitative content analysis in research*. New York, NY: Routledge.

Ruggles, Steven. 2014. Big microdata for population research. *Demography* 51 (1): 287–97.

Shah, Dhavan V., Joseph Cappella, and W. Russell Neuman. 2015. Big data, digital media, and computational social science: Possibilities and perils. *The ANNALS of the American Academy of Political and Social Science* (this volume).

Sjøvaag, Helle, Hallvard Moe, and Eirik Stavelin. 2012. Public service news on the web: A large-scale content analysis of the Norwegian Broadcasting Corporation's online news. *Journalism Studies* 13 (1): 90–106.

Sjøvaag, Helle, and Eirik Stavelin. 2012. Web media and the quantitative content analysis: Methodological challenges in measuring online news content. *Convergence: The International Journal of Research into New Media Technologies* 18 (2): 215–29.

Stone, Philip J., Robert F. Bales, J. Zvi Namenwirth, and Daniel M. Ogilvie. 1962. The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science* 7 (4): 484–98.

Thelwall, Mike, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology* 61 (12): 2544–58.

Vliegenthart, Rens, and Conny Roggeband. 2007. Framing immigration and integration. *International Communication Gazette* 69 (3): 295–319.

Weller, Katrin, Axel Bruns, Jean Burgess, Merja Marht, and Cornelius Puschmann, eds. 2014. *Twitter and society*. New York, NY: Peter Lang.