Empirical/Policy

Social Currents
2019, Vol. 6(4) 307–318

DOI: 10.1177/2329496519846505
journals.sagepub.com/home/scu
SAGE

Structural Topic Modeling For Social Scientists: A Brief Case Study with Social Movement Studies Literature, 2005–2017

Nathan C. Lindstedt[1]

Abstract

Sociologists frequently make use of language as data in their research using methodologies including open-ended surveys, in-depth interviews, and content analyses. Unfortunately, the ability of researchers to analyze the growing amount of these data declines as the costs and time associated with the research process increases. Topic modeling is a computer-assisted technique that can help social scientists to address these data challenges. Despite the central role of language in sociological research, to date, the field has largely overlooked the promise of automated text analysis in favor of more familiar and more traditional methods. This article provides an overview of a topic modeling framework especially suited for social scientific research. By way of a case study using abstracts from social movement studies literature, a short tutorial from data preparation through data analysis is given for the method of structural topic modeling. This example demonstrates how text analytics can be applied to research in sociology and encourages academics to consider such methods not merely as novel tools, but as useful supplements that can work beside and enhance existing methodologies.

Keywords

topic modeling, methodology, collective behavior and social movements

The *sine qua non* of sociology has been the attempt to understand and explain the complexities of group conduct. In doing so, the field relies heavily on information communicated through spoken or written language. Research methodologies such as open-ended surveys, in-depth interviews, and content analyses are, fundamentally, a set of practices established for the orderly examination of some underlying linguistic data. Given the importance of language as data for much of their research, sociologists today generally lag behind other disciplines in innovating analyses of language and its meaning, just as the amount of textual data readily available for its study is quickly growing in volume. Some scholarship has recognized the need for new tools that are better able to handle the increasing abundance of data as academics seek to extend and build theories of social life (e.g., Evans and Aceves 2016). Other scholarship has gone so far as to

[1]Washington State University, Pullman, USA

Corresponding Author:
Nathan C. Lindstedt, Department of Sociology, Washington State University, PO Box 644020, Pullman, WA 99164, USA.
Email: nathan.lindstedt@wsu.edu

proclaim that empirical sociology is nearing a state of "crisis" because of its inattentiveness toward these methodological advances (e.g., Savage and Burrows 2007). Several authors have recommended combining the study of language and its meaning with computer-based procedures for automated analyses of textual data (e.g., Bail 2014; Lindstedt 2018). Responding to those calls, this article provides an introduction to structural topic modeling for social scientists looking to expand their methodological repertoire to meet the coming data challenges.[1]

Latent Dirichlet Allocation (LDA) Meets Social Science

Topic modeling is a group of inductive techniques used to discover hidden topics contained in textual data. First developed by David M. Blei, Andrew Y. Ng, and Michael I. Jordan (2003), mixed-membership topic modeling, also known as LDA, is an unsupervised method for identifying key topics within a collection of documents. Beneath LDA rests a generative statistical model that assumes that the observed documents are produced from a mixture of latent topics.[2] These unobserved topics, whose number is defined by the researcher beforehand, then generate associated words based on their probability distributions. Thus, given a set of documents and a number of topics, LDA attempts to identify what combination of the unknown topics could generate those documents.

Extending the LDA framework in promising directions for social scientific research is the structural topic model (STM). The STM approach builds upon the standard topic model, but it has some specific advantages that are appealing to social scientists: it can accommodate supplemental information in the form of covariates that could reveal important aspects of how topics are discussed *or* that could help to describe the frequency with which topics are discussed (Roberts et al. 2013). In the former case, covariates concerning *topical content* can answer research questions regarding differences in the language used to discuss topics (e.g., political ideology, geographical location, etc.), while in the latter case, covariates concerning *topical prevalence* can answer research questions regarding differences in the frequency with which topics are discussed (e.g., dates, authorship, etc.). The added benefits of the ability to incorporate model covariates to address some of these research questions are presented in the following case study.

Case Study Using Social Movement Studies Literature

The study of collective behavior and social movements has long been a core interest of sociology.[3] But although social movement studies represent a relatively established domain of sociological analysis, that does not mean that subject interests have become fixed. Owing to that, the area provides a suitable case for exploring what, if any, significant changes have occurred in this ostensibly developed subfield over time.

Throughout the years, the rise and fall of research programs has led to revisions in our understanding of social movements time and again. Among these broader changes are the progression from "classical" theories of collective behavior (e.g., Smelser 1962), wherein social movements emerge as a response to system strain, to more contemporary theories of fields (e.g., Fligstein and McAdam 2012), wherein contentious action is contingent on a competitive process between incumbents and challengers. Along with these larger programmatic changes, the topical interests of researchers have changed as well. For one, current events have altered the setting in which scholars produce knowledge. Societal moments, such as the appearance of the #BlackLivesMatter and the #MeToo movements, may direct researcher attention toward certain lines of inquiry, which could ultimately shift the direction of future academic work (Moody and Light 2006).

To gain a better grasp on the current standing of knowledge produced in the subfield, and its potential impact, it is important to take stock of what knowledge has been produced in the past. That is, to understand where you are,

and where you might be headed, it is good to know where you have been. This brief case study provides a gentle introduction to text analytics as applied to the case of contemporary social movement studies. It aims to give social scientists the practical knowhow of a quantitative technique that is suited to examining these and other sorts of thematic trends.[4]

The process presented results in 24 key topics from social movement scholarship being identified in the abstracts of 11 top national and regional sociology journals and tracks changes in the prevalence and impact of those topics during the 2005–2017 period. In this demonstration, the document set is made up of abstracts from academic journals, but they could easily be drawn from other sources such as newspaper articles (e.g., DiMaggio, Nag, and Blei 2013), open-ended survey responses (e.g., Tingley 2017), speeches (e.g., Light and Cunningham 2016), public comments (e.g., Levy and Franklin 2014), or several combined sources (e.g., Farrell 2016a) to address a variety of research questions.[5] To complete this analysis, the stm R package developed by Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley (2018) is used.

While other recent examples of scholarship in sociology using STMs exist (e.g., Almquist and Bagozzi 2017; Bohr and Dunlap 2017), none discuss many of the practical decisions that need to be made to accomplish such research.[6] This work addresses that gap by demystifying the procedure used to arrive at readily interpretable and analytically useful results through a summary of the leading recommendations made to date by practitioners of structural topic modeling. Using data from an online citation index and open-source software, this article is a primer on data preparation, model selection, estimation, diagnostic evaluation, and data analysis within the STM framework.

Data Collection and Methods

The data for this analysis were collected using the Web of Science's Social Science Citation Index (SSCI). Included were abstracts from the following journals: *American Journal of Sociology, American Sociological Review, Mobilization, Social Forces, Social Problems, Social Science Quarterly, Sociological Forum, Sociological Inquiry, Sociological Perspectives, Sociological Quarterly*, and *Sociological Spectrum* (see Table 1). The majority of journals were selected for inclusion because they represent the top general interest sociology journals on a national and regional level. Also included was the specialty journal *Mobilization* for its relevance to the subarea of study. Other journals were omitted because either they are relatively recent additions to the field, such as *Social Currents*, or they are not contained in the SSCI, such as *Research in Social Movements, Conflicts and Change*.

Table 1. Descriptive Statistics for Sociology Journal Abstracts.

Academic journal	Frequency	%
American Journal of Sociology	51	7.22
American Sociological Review	64	9.07
Mobilization	299	42.35
Social Forces	54	7.65
Social Problems	59	8.36
Social Science Quarterly	8	1.13
Sociological Forum	55	7.79
Sociological Inquiry	26	3.68
Sociological Perspectives	38	5.38
Sociological Quarterly	39	5.52
Sociological Spectrum	13	1.84
Total	706	100

Aside from *Mobilization*, articles were only considered for inclusion if they contained the keyword search terms "social movement" or "social movements." This choice greatly narrowed the scope of the articles included, and similarly restricted keyword searches may not provide the best results for researchers given their specific areas of interest. During this early stage of the research process, significant time should be invested by the researcher into considering what documents should and should not be included in their dataset. In this case, trials with closely related terms such as "collective behavior," "collective action," "activism," and "protest" yielded results with substantial bounding issues. In other words,

combinations of these successive terms returned search results with a large number of articles from outside of the defined area of interest, a potential source of troublesome bias.

Jeremiah Bohr and Riley Dunlap (2017) noted that using online citation databases, such as the SSCI, for data collection purposes often involves a balancing act between obtaining a clearly bounded dataset and capturing each pertinent article due to the ambiguities of language. Recent work has suggested the use of computer-assisted techniques for improved keyword selection and cautions that "researchers usually pick keywords in *ad hoc* ways that are far from optimal and usually biased" (King, Lam, and Roberts 2017:1). Unfortunately, an in-depth summary of this approach for keyword discovery from unstructured text is beyond the scope of this article. Generally speaking, however, keyword searches that confine the set of documents to a well-defined area of interest increases the inferential performance of topic modeling and will serve to limit the number of false positives, although the likelihood of overlooking relevant documents remains. In total, this data collection procedure generated 706 abstracts.

Besides the textual data from the abstracts, the SSCI includes additional data for each article entry including the year of publication, total times cited, and periodic usage counts. This so-called metadata can then be used as covariates inside the STM framework. For the purpose of demonstrating how to integrate an additional topical prevalence covariate into a topic model, a measure of impact using average citations per year is constructed from the included metadata by dividing the total times cited by the number of years since publication. Beyond the benefit of being able to account for model covariates, the stm R package also allows for the estimation of correlations between topics. This attribute is desirable because, as noted before, in topic modeling, documents are made up of mixtures of topics. Researchers, therefore, can use structural topic modeling to observe which topics correlate strongly with one another on a document level. Graphs of these correlations are then drawn using the Fruchterman-Reingold algorithm found in the igraph R package (Csardi and Nepusz 2006).

The results of a 24-topic solution, and their calculated correlations, are presented in the next section. Prior to their presentation, however, a discussion of the pitfalls of topic modeling in general and model selection and model diagnostics in particular is warranted. As scholars Justin Grimmer and Brandon M. Stewart (2013) have warned, while automated methods offer substantial promise for reducing the costs and time typically associated with more traditional analyses of textual data, they are imperfect substitutes for human interpretations and domain-specific expertise.[7] Simply put, automated methods are not a replacement for, but are a supplement to, the abilities of researchers in a multistage research process.[8,9] One drawback of topic modeling techniques noted by academics is that there are no straightforward means for selecting a number of topics that produce both readily interpretable and analytically useful results (e.g., Farrell 2016b; Roberts et al. 2014). The crux of the matter is that within LDA, it is up to the researcher to choose the number of topics in advance. Indeed, because automated text analyses are based on incorrect models of language, there is not a single correct answer for selecting the number of topics (see Grimmer and Stewart 2013). The assumption of the probabilistic "bag of words" model of language used for the extraction of topics from textual data is that topics can be inferred from the co-occurrence of terms among documents wherein the word ordering does not inform the analysis. Moreover, while there exist a variety of means for assessing model fit, strict adherence to these diagnostics can lead to ambiguous results. In part, the reason for this is that when topic models are used for prediction, those models that are best able to predict out-of-sample documents often do not coincide with human judgments. Consequently, model selection based upon optimizing held-out likelihood measures can result in less than insightful outcomes—a phenomenon known as the prediction-interpretability trade-off (Chang et al. 2009; Wesslen 2018). Those caveats aside, there are other useful guidelines

Table 2. Structural Topic Model Diagnostic Table.

No. of topics (*K*)	Exclusivity	Semantic coherence	Held-out likelihood	Residual
5	9.310	−71.253	−5.669	1.282
10	9.537	−80.076	−5.645	1.141
15	9.433	−79.853	−5.640	1.068
20	9.559	−83.056	−5.607	1.017
25	9.632	−87.440	−5.634	0.975
30	9.686	−88.139	−5.621	0.945
35	9.705	−90.539	−5.613	0.920
40	9.729	−90.367	−5.583	0.905
45	9.775	−91.677	−5.587	0.901
50	9.766	−93.159	−5.593	0.891

and diagnostics that, when followed, can coach researchers toward good starting points for their analyses.

Thankfully, the authors of the stm R package provide some direction on how to proceed with the thorny issue of selecting the number of topics, although they concede that there are no established means for obtaining consistent results across different sets of documents. A preliminary, more subjective, recommendation is related to the relationship held between document length, document focus, and the performance of LDA models. For shorter, focused corpora (i.e., those ranging from a few hundred to a few thousand documents in size), an initial choice between five and 50 topics is best, whereas for larger, unfocused corpora (i.e., those ranging from tens of thousands to hundreds of thousands of documents in size or larger), previous research has found that between 60 and 100 topics are best (Roberts et al. 2018).

A secondary, less subjective, recommendation involves examining diagnostic tables and plots of *semantic coherence* and *exclusivity* calculations. Semantic coherence is a measure of the probability for a set of topic words to co-occur within the same document. Exclusivity is a measure of the probability for a word to fall primarily within the top rankings of a single topic. Model selection for the number of topics is made along the semantic coherence-exclusivity "frontier" where no model is dominated by either metric (Roberts et al. 2014). See Table 2 for the diagnostic output and Figure 1 for the semantic coherence-exclusivity plot of models across the preliminary suggested range of topic numbers. More often than not, researchers will want to narrow down their model selection based on it having desirable properties along both the semantic coherence and exclusivity dimensions. That is, they will begin by choosing candidate models with solutions that place them nearer the upper right-hand quadrant of the plot (Roberts, Stewart, and Tingley forthcoming). In this example, a model with around 20 topics seemingly offers the best balance between each metric. Conversely, a model with around 15 topics, which rates higher on semantic coherence but lower on exclusivity, is likely not as suitable a solution in terms of its interpretability.

Given the difficulties associated with model selection and the trade-off between predictive and interpretable models, the ultimate responsibility for model selection rests with the researcher and their informed judgment. Therefore, it is on the researcher to "validate, validate, validate" their results (Grimmer and Stewart 2013:5). This process can be done in a number of ways, but the most useful means of validation in the stm R package is its built-in function that provides a list of the most representative documents for a particular topic. Another method is to evaluate the topic quality by assessing the word clusters that make up each topic (Roberts et al. 2014). Additional built-in functions are provided that display closely associated words within each topic, including those words with the highest probability to be found within each topic *or* words

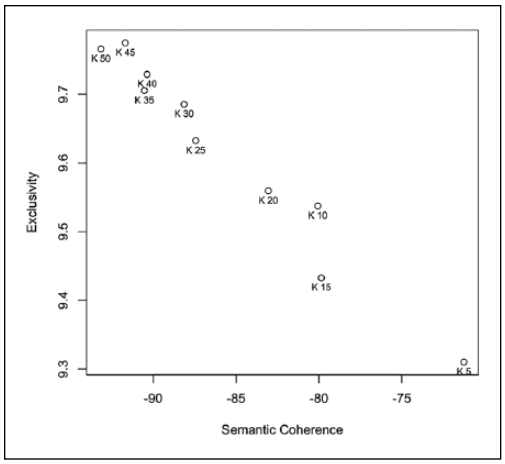

Figure 1. Structural topic model semantic coherence-exclusivity plot.

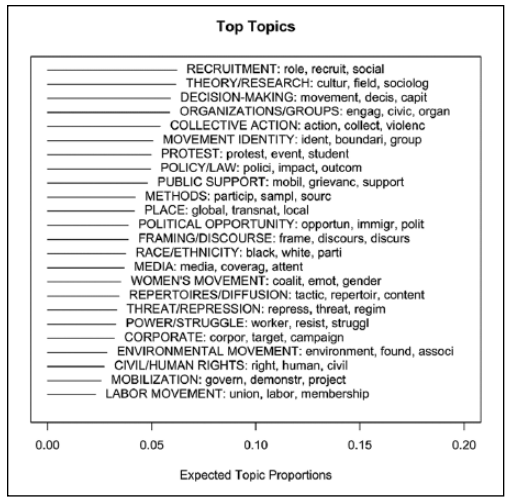

Figure 2. Labels for a 24-topic solution.

weighted by their frequency and exclusiveness within each topic (Roberts et al. forthcoming). Based on the initial number of around 20 topics, which was suggested by the semantic coherence-exclusivity diagnostics of the candidate models, these combined built-in functions were used to qualitatively assess the results of each possible model in the 20 to 25 topic range to arrive at the final number of topics.

Results

Figure 2 presents the results of the 24-topic solution. At the top end of the scale, and representing around 6 percent of the documents, the most prevalent topic was "recruitment." At the bottom end of the scale, and representing around 2 percent of the documents, the least prevalent topic was "labor movement." For social movement scholars, it is probably not unexpected that topics concerning organizations, movement identity, and political opportunity are featured prominently in this ranking because these topics speak to some of the major theoretical outlooks in the subfield. But the toolkit offered by the stm R package allows for researchers to go a step further by identifying prototype documents (i.e., those documents with the highest proportion of words devoted to a particular topic) inside of each topic area. For instance, when the recruitment topic is queried for a prototype document, Whittiers's (2014) article on rethinking coalitions in the anti-pornography movement is returned. When asked for more documents within the recruitment topic, White's (2007) methodological piece on recruitment into Irish Republicanism, Munson's (2010) examination of how changes in social network configurations explain college conservative mobilization, and Crossley's (2015) look at online feminist forums expanding recruitment bases are returned. Clearly, these are all documents pertaining to issues of recruitment, spillover, and coalition building.

But observing topic proportions and their representative documents does not tell us how these topics have been trending over the years. To get a more complete picture of the knowledge produced in the subfield as a function of time, the year of publication can be used as a topical prevalence covariate and the resulting estimates of the changes in topic proportions can then be plotted. For the purpose of demonstration, the "women's movement" and "race/ethnicity" topics were selected because they highlight a particularly drastic reversal in topic proportion during the observation period (see Figure 3). Over the 2005–2017 period, the women's movement topic trended downward from representing around 4.5 percent to 2.4 percent of the reported topic proportions. By contrast, the race/ethnicity topic trended upward from around 2.2 percent to 5.1 percent of the reported topic proportions.

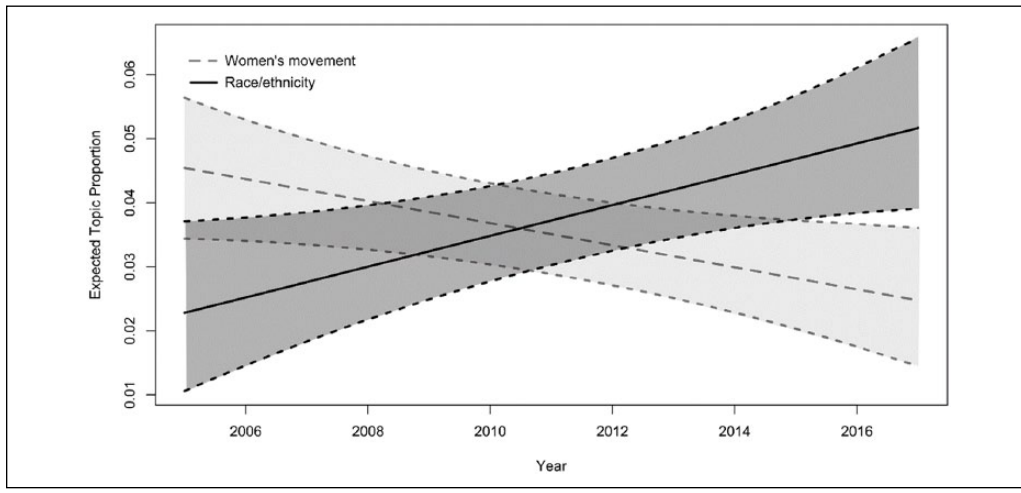

Figure 3. Topic prevalence for "women's movement" and "race/ethnicity," by year (with 95% confidence intervals).

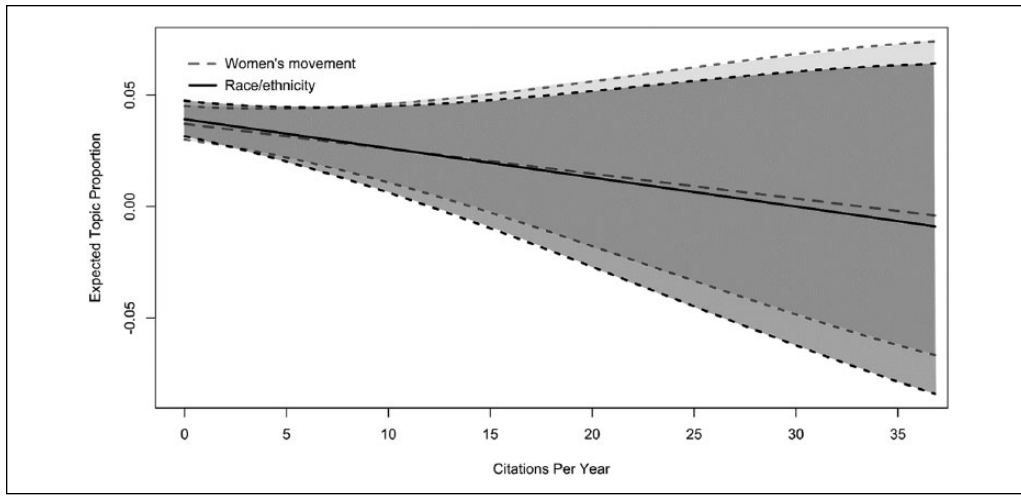

Figure 4. Topic prevalence for "women's movement" and "race/ethnicity," by citations per year (with 95% confidence intervals).

To close this discussion on some of the additional data analysis options that are available through the use of covariates, we will revisit the variable of average citations per year. Although there were clearly changes in the prevalence of the women's movement and race/ethnicity topics over the years, that alone reveals nothing about the prevalence of those topics across different levels of impact. To do that, the measure of average citations per year can also be introduced as a model covariate. For the entire set of documents, values range from no citations per year on the low end to 36.82 citations per year on the high end—with a mean value of 2.04 citations per year, a median value of 1.17 citations per year, and a standard deviation of 3.08 citations per year. Note that despite the reversal in their frequency over time, each topic has a similar impact (see Figure 4). That is, within two standard deviations of the mean value for average citations per year, the frequency of the "women's movement" and "race/ethnicity" topics remain at comparable levels. By comparison, Figure 5

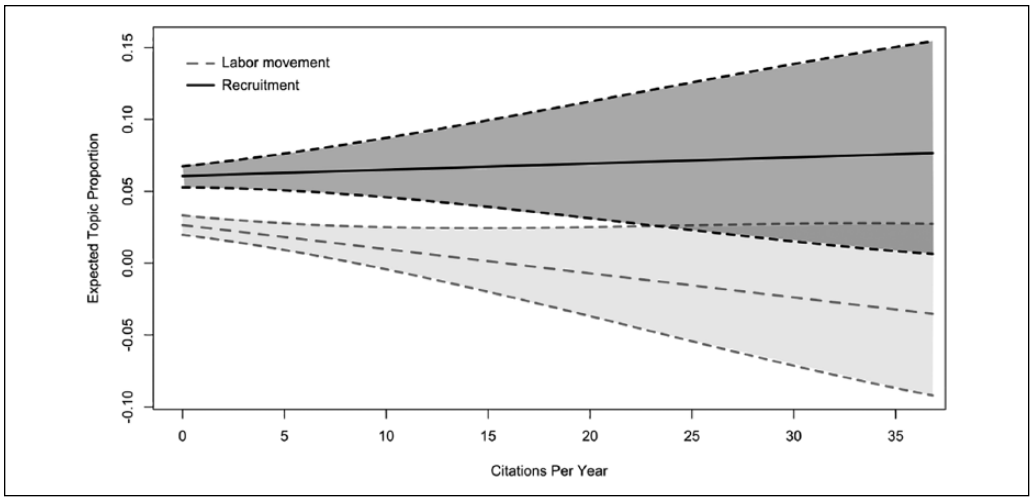

Figure 5. Topic prevalence for "labor movement" and "recruitment," by citations per year (with 95% confidence intervals).

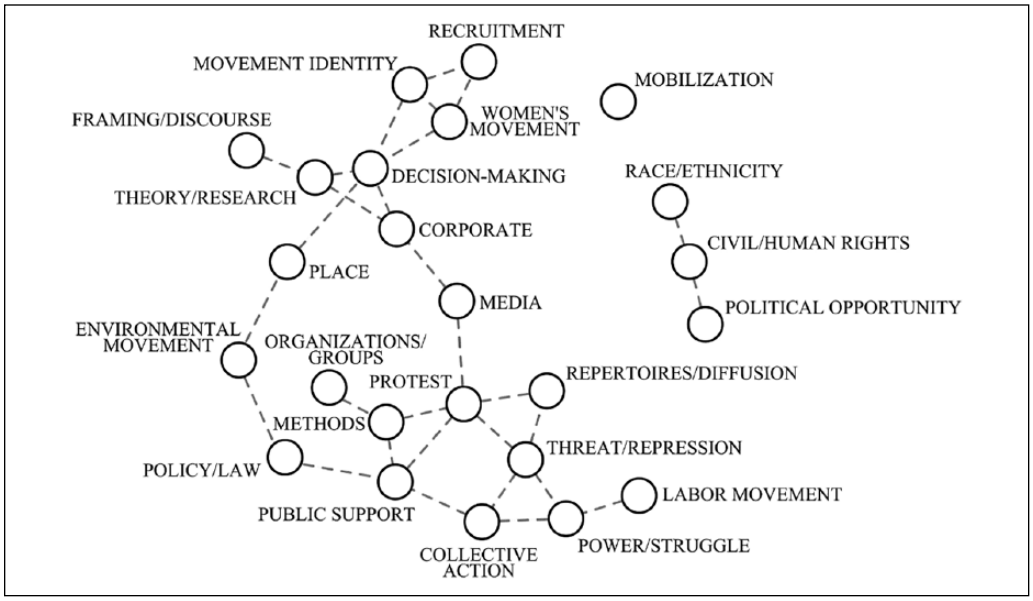

Figure 6. Topic correlation graph (with tie strength $>$ .01).

presents an example where the influence of impact on topic prevalence varies to a larger degree. Among articles within two standard deviations of mean value, documents on the topic of "recruitment" make up a significantly far greater share than those on the topic of "labor movement."

Figure 6 presents the graph of topic correlations on a document level. Notice that of the paired topics of interest, neither "women's movement" and "race/ethnicity" nor "labor movement" and "recruitment" are significantly correlated with tie strengths greater than the .01 level. Positive correlations above this threshold indicate that the paired topics are likely to be discussed within the same document. Therefore, it appears these pairs do not represent closely related topics. But astute

social movement scholars might spot other important features of the subfield in the graph. These are signified by the dyads and triads formed between topics such as "women's movement," "recruitment," and "movement identity"; "women's movement," "decision-making," and "movement identity"; "race/ethnicity" and "civil/human rights"; and "labor movement" and "power/struggle." One substantive interpretation of these connections between correlated topics is that the presence of such ties represents proximate areas of study, which have a greater propensity for influencing one another, whereas the absence of such ties represents distal areas of study, which have a lower propensity for influencing one another (Fligstein and McAdam 2012).

By now, some of the potential uses of these tools for researchers working with a variety of textual data sources should be discernible. The STM extension of topic modeling not only permits a researcher to accomplish those things that can already be done from within the standard LDA framework, namely, the identification of latent topics, it also lets them get at other questions that are of particular relevance to social scientists concerning how textual data are structured using additional covariates. With these benefits in mind, the results illustrate the usefulness of this technique for social scientific research. Although not an exhaustive resource, this article demonstrates a few of the possible uses of structural topic modeling for the analysis of textual data. Other work might consider how covariates of topical content, such as ideological outlook or geographical region, affect what language is used to discuss these topics. While there are important limitations to what structural topic modeling can accomplish, it also opens up additional lines of inquiry for those researchers willing to engage with the method.

Conclusion

Sociologists routinely make use of language as data. However, they have begun to lag behind other disciplines in pioneering approaches that are able to make use of the ever-growing availability of such data. This article presents an instructive tutorial for researchers wanting to expand their methodological repertoire to be better positioned to meet these coming data challenges. Through the use of structural topic modeling, academics can summarize the content and examine the structure of topics within large collections of documents in ways that are prohibitive to more traditional analyses alone. Indeed, the STM framework can be used in conjunction with other methodologies (including open-ended surveys, in-depth interviews, and content analyses) to achieve analyses previously too costly or too time-consuming to pursue, *while* enabling scholars to gain leverage over important research questions. That is, it can be used in a multistage research process to expedite the identification of the principal topics within sizable textual datasets to motivate subsequent closer qualitative analyses. In the included case study, the added benefits of the STM framework, such as the use of model covariates, display a few of the prospective uses of this technique in future scholarship. In sum, this article provides researchers with an introduction to structural topic modeling and encourages them to use the tools of text analytics to help them innovate beyond more familiar methodological territory.

Acknowledgments

The author would like to thank Erik W. Johnson, the anonymous reviewers, and the editors for their helpful comments and suggestions on earlier drafts of this article.

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) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. The R script and the data used to complete the automated text analysis described in this article are provided as supplemental files online so that the reader can follow along and replicate the findings as detailed.

2. A "generative" model is a statistical model that generates the unobserved inputs given the observed outputs. This is in contrast to a "discriminative" model, which infers the unobserved outputs given the observed inputs. Latent Dirichlet allocation (LDA) is an example of a generative model whereas logistic regression is an example of a discriminative model. For more detail on the generative process used, Blei's (2012) description of LDA using plate notation provides a working basis for understanding the general principles at work.
3. In the opening of his book *Social Conflict and Social Movements*, Oberschall (1973) observed, "To write about the antecedents of macrotheories of social conflict and social movements would be tantamount to writing the history of sociological thinking itself, starting with Marx and Tocqueville" (p. 1).
4. Ignatow and Mihalcea (2017, 2018) have released a series of books that are also aimed at providing social scientists with an introduction to the text analytics methods described in this article.
5. DiMaggio, Nag, and Blei (2013) used newspaper articles to identify the primary themes that framed media coverage of government grants supporting the arts between 1986 and 1997. Tingley (2017) used open-ended survey responses to explore the logics articulated by individuals under conditions of declining power or rising power in international relations. Light and Cunningham (2016) used speeches of Nobel laureates in the area of peace to uncover themes and then used the results to drive their qualitative analysis. Levy and Franklin (2014) used public comments on proposed trucking regulations to understand differences in how individual and organizational stakeholders framed a policy debate over the electronic monitoring of truckers' work hours. Farrell (2016a) used several combined sources, including written and verbal texts, to show that corporate funding of climate change countermovement organizations influenced the thematic content of their polarizing discourse.
6. Other scholarship, such as Mohr and Bogdanov (2013), provides a nontechnical introduction to topic modeling, but not the structural topic modeling variant, which is the focus of this article.
7. To help mitigate this issue, studies have shown that the greater the number of documents that concentrate on relatively few topics, and the longer those focused documents are, the better the LDA will perform (Tang et al. 2014). Solitary volumes covering a range of topics, such as a single book, or particularly short documents, such as tweets, provide less than ideal datasets for topic modeling. In these cases, alternative techniques like dividing books into multiple cohesive documents (e.g., Mimno and McCallum 2007) or aggregating tweets based on shared attributes (e.g., Hong and Davison 2010) have been used by researchers.
8. Nelson (2017) has proposed a framework for computational grounded theory that can be applied to textual data including open-ended surveys, in-depth interviews, and content analyses. In the first stage, computer-based text analysis techniques help researchers to identify previously unaccounted latent patterns while staying grounded in the data. In the second stage, researchers work on a subset of the texts to confirm the credibility of the identified latent patterns, interpret the results, and adjust the computational models to enhance their interpretation. In the third stage, researchers test whether the results from the earlier stages are generalizable and are the final check for the reliability of the computational grounded theory process.
9. For example, Valdez, Pickett, and Goodson (2018) contended that topic modeling can be used either for coding purposes or as a means to confirm the already generated codes. In the first case, topic modeling is employed by researchers to discover topics and investigate how they are structured. In the second case, topic modeling is employed by researchers subsequent to their coding of qualitative data to serve as an additional reliability check of the determined codes.

Supplemental Material

Supplemental material for this article is available online.

ORCID iD

Nathan C. Lindstedt https://orcid.org/0000-0002-5263-5687

References

Almquist, Zack W. and Benjamin E. Bagozzi. 2017. "Using Radical Environmentalist Texts to Uncover Network Structure and Network Features." *Sociological Methods & Research*.

Published electronically November 16. doi: 10.1177/0049124117729696.

Bail, Christopher A. 2014. "The Cultural Environment: Measuring Culture with Big Data." *Theory and Society* 43(3–4):465–82.

Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55(4):77–84.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3:993–1022.

Bohr, Jeremiah and Riley E. Dunlap. 2017. "Key Topics in Environmental Sociology, 1990–2014: Results from a Computational Text Analysis." *Environmental Sociology* 4(2):181–95.

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." Pp. 288–96 in *Advances in Neural Information Processing Systems*, edited by Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta. 22nd ed. Red Hook, NY: Curran Associates.

Crossley, Alison Dahl. 2015. "Facebook Feminism: Social Media, Blogs, and New Technologies of Contemporary U.S. Feminism." *Mobilization: An International Quarterly* 20(2):253–68.

Csardi, Gabor and Tamas Nepusz. 2006. "The Igraph Software Package for Complex Network Research." *InterJournal Complex Systems* 1695:1–9.

DiMaggio, Paul, Manish Nag, and David Blei. 2013. "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding." *Poetics* 41(6):570–606.

Evans, James A. and Pedro Aceves. 2016. "Machine Translation: Mining Text for Social Theory." *Annual Review of Sociology* 42(1):21–50.

Farrell, Justin. 2016a. "Corporate Funding and Ideological Polarization about Climate Change." *Proceedings of the National Academy of Sciences of the United States of America* 113(1):92–97.

Farrell, Justin. 2016b. "Corporate Funding and Ideological Polarization about Climate Change." Supplementary Information. Retrieved September 4, 2018 (http://www.pnas.org/content/pnas/suppl/2015/11/18/1509433112.DCSupplemental/pnas.1509433112.sapp.pdf).

Fligstein, Neil and Doug McAdam. 2012. *A Theory of Fields*. New York: Oxford University Press.

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.

Hong, Liangjie and Brian Davison. 2010. "Empirical Study of Topic Modeling in Twitter." Pp. 80–88 in *Proceedings of the First Workshop on Social Media Analytics, SOMA'10*. New York: ACM.

Ignatow, Gabe and Rada Mihalcea. 2017. *Text Mining: A Guidebook for the Social Sciences*. Los Angeles, CA: Sage Publications.

Ignatow, Gabe and Rada Mihalcea. 2018. *An Introduction to Text Mining: Research Design, Data Collection, and Analysis*. Thousand Oaks, CA: Sage Publications.

King, Gary, Patrick Lam, and Margaret E. Roberts. 2017. "Computer-Assisted Keyword and Document Set Discovery from Unstructured Text." *American Journal of Political Science* 61(4):971–88.

Levy, Karen E. C. and Michael Franklin. 2014. "Driving Regulation: Using Topic Models to Examine Political Contention in the U.S. Trucking Industry." *Social Science Computer Review* 32(2):182–94.

Light, Ryan and Jeanine Cunningham. 2016. "Oracles of Peace: Topic Modeling, Cultural Opportunity, and the Nobel Peace Prize, 1902–2012." *Mobilization: An International Quarterly* 21(1):43–64.

Lindstedt, Nathan. 2018. "Shifting Frames: Collective Action Framing from a Dialogic and Relational Perspective." *Sociological Compass* 12(1):1–12.

Mimno, David and Andrew Mccallum. 2007. "Organizing the OCA: Learning Faceted Subjects from a Library of Digital Books." Pp. 376–385 in Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, *JCDL '07*. Vancouver, British Columbia, Canada: ACM.

Mohr, John W. and Petko Bogdanov. 2013. "Introduction—Topic Models: What They Are and Why They Matter." *Poetics* 41(6):545–69.

Moody, James and Ryan Light. 2006. "A View from Above: The Evolving Sociological Landscape." *American Sociologist* 37(2):67–86.

Munson, Ziad. 2010. "Mobilizing on Campus: Conservative Movements and Today's College Students." *Sociological Forum* 25(4):769–86.

Nelson, Laura K. 2017. "Computational Grounded Theory: A Methodological Framework." *Sociological Methods & Research*. Published electronically November 21. doi:10.1177/0049124117729703.

Oberschall, Anthony. 1973. *Social Conflict and Social Movements*. Englewood Cliffs, NJ: Prentice-Hall.

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2018. "Package 'stm.'" R

Package Version 1.3.3. Retrieved September 4, 2018 (https://cran.r-project.org/web/packages/stm/stm.pdf).

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. Forthcoming. "stm: R Package for Structural Topic Models." *Journal of Statistical Software* 1: 12. Retrieved September 4, 2018 (https://cran.r-project.org/web/packages/stm/vignettes/stmVignette.pdf).

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, and Edoardo M. Airoldi. 2013. "The Structural Topic Model and Applied Social Science." Presented at the Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation, December 24, Cambridge, MA.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58(4): 1064–82.

Savage, Mike and Roger Burrows. 2007. "The Coming Crisis of Empirical Sociology." *Sociology* 41(5):885–99.

Smelser, Neil. 1962. *Theory of Collective Behavior*. New York: Free Press.

Tang, Jian, Zhaoshi Meng, XuanLong Nguyen, Qiaozhu Mei, and Ming Zhang. 2014. "Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis." Pp. 190–8 in *Proceedings of the 31st International Conference on International Conference on Machine Learning—Volume 32, ICML'14*, edited by E. P. Xing and T. Jebara. Beijing, China: Journal of Machine Learning Research.

Tingley, Dustin. 2017. "Rising Power on the Mind." *International Organization* 21:S165–S88.

Valdez, Danny, Andrew C. Pickett, and Patricia Goodson. 2018. "Topic Modeling: Latent Semantic Analysis for the Social Sciences." *Social Science Quarterly* 99(5):1665–79.

Wesslen, Ryan. 2018. "Computer-Assisted Text Analysis for Social Science: Topic Models and Beyond." Charlotte, NC: University of North Carolina at Charlotte.

White, Robert. 2007. "'I'm Not Too Sure What I Told You the Last Time': Methodological Notes on Accounts from High-Risk Activists in the Irish Republican Movement." *Mobilization: An International Quarterly* 12(3):287–305.

Whittier, Nancy. 2014. "Rethinking Coalitions: Anti-Pornography Feminists, Conservatives, and Relationships between Collaborative Adversarial Movements." *Social Problems* 61(2):175–93.