Large-scale Computerized Text Analysis in Political Science: Opportunities and Challenges

John Wilkerson and Andreu Casas

Annual Review of Political Science YYYY.
AA:1–18

This article's doi:
10.1146/((XXXXXXXXXXX))

Keywords

text as data, automatic coding, machine learning, computational social sciences

Abstract

Text has always been an important data source in political science. What has changed in recent years is the feasibility of investigating large amounts of text quantitatively. The internet provides political scientists with more data than their mentors could have imagined, and the research community is providing accessible text analysis software packages, along with training and support. As a result, text as data research is beginning to mainstream in political science. Scholars are tapping new data sources, they are employing more diverse methods, and they are becoming critical consumers of findings based on those methods. In this article, we first introduce readers to the subject by describing the four stages of a typical text as data project. We then review recent political science applications, and explore one important methodological challenge - topic model instability - in greater detail.

Contents

1. Introduction

Words are an integral part of politics. Officials and citizens use words to express opinions, make proposals, and defend their actions. Laws and regulations are also largely codified in words. Political scientists have always been interested in words but a revolution has occurred that is creating unprecedented research opportunities for political scientists (Monroe and Schrodt 2008; Cardie and Wilkerson 2008; Alvarez 2016). The internet is providing an avalanche of data related to politics. For example, all departments and agencies of the U.S. Federal government must now post their public records on the internet and many other governments adhere to similar practices. Most major newspapers offer on-line access to their archives. Project Gutenberg and Google Books offer free access to the complete texts of millions of books. Social media sites such as Twitter and Facebook encourage researchers to use their data. The Internet Archive offers archival information about millions of government web pages dating back to 1996.

The research community has responded to this surfeit of data by developing accessible open source text analysis libraries in R, Python, and other programming languages (e.g. Munzert et al. 2014). As a result, political scientists have access to a variety of tools and methods for efficiently analyzing large volumes of text.[1] The combination of so many untapped research opportunities and accessible tools and training make this an excellent time for specialists in all areas to invest in text. Legislative scholars can now systematically investigate floor speeches, constituent communications, revisions to laws and regulations, and much more. International Relations scholars can systematically compare final treaties or agreements to hundreds of proposals made at earlier stages. Political theorists can explore political thought by searching across centuries of published works.

[1] One new R package, `quanteda`, may be especially useful for scholars exploring text as data methods for the first time.

This new found ability to investigate text computationally (as well as many other innovative data sources such as images and sound) will transform political science research as scholars become more adept at exploiting the available opportunities. Because not all readers may be familiar with text-as-data research, we first provide an overview of the four stages of a typical project. This overview provides context for appreciating recent developments and methodological challenges. We then review recent political science applications, and explore one important methodological challenge - topic model instability - in greater detail.

2. Four Stages of a Text as Data project

Text as data methods expand research opportunities for political scientists in two ways. First, they leverage the power of computing to make ambitious data collection tasks feasible. Second, they offer a growing number of options for analyzing large volumes of text quantitatively. A typical text as data project proceeds through four stages. Text must be obtained, converted to quantitative data, analyzed and validated.

2.1. Obtaining Text

The first stage of a project usually entails downloading digitized content. For many projects, this is now a fairly minor step. However, it is probably wise to investigate what will be required before committing to a project. Some content providers offer APIs (Application User Interfaces) that enable users to "request" selected content from an underlying structured database using a single line of code. APIs are ideal when they include options that serve the needs of a project. Later in this paper we show how a Sunlight Foundation API can be used to select only member's statements from volumes of the Congressional Record. Examples include the multiple APIs offered by the New York Times (e.g. Article Search API, Congress API, etc.), the Sunlight Foundation (e.g. Open States, Capitol Words, etc.) and prominent social media sites (Twitter, Facebook etc.).

If an API is not available, the next best option in terms of ease of use are documents that are similarly formatted. Identical formatting makes it possible to write a single script to extract more specific content from many documents at once, such as the thousands of congressional bill texts available through the Government Printing Office. Almost all documents contain hidden formatting language that may also be helpful for systematically extracting more specific content. The look and feel of a web page comes from embedded .html or .xml tags. These tags may do little more than format the visible text, but they can also be used to isolate desired content (see for example the @unitedstates project (https://theunitedstates.io/)). Other types of documents (.doc, . docx., txt, .pdf) also contain hidden formatting that may provide unique markers to facilitate parsing. Even consistently formatted text can be helpful. In transcripts of Federal Reserve Board meetings, only the speaker's name is printed in all capital letters ("MS. YELLEN.") and can be used to easily parse transcripts by speaker statement.

The most challenging scraping projects are those that draw content from diverse sources with inconsistent formatting. For example, extracting the same content from many different candidate websites is challenging because each website has a different structure. One option is to write multiple scripts. The OpenStates project (openstates.org) recruited volunteer programmers to write scores of scripts to extract information about legislative bills for di?erent state government websites. For less ambitious projects, crowdsourcing may be more practical. Sites such as Mechanical Turk or Crowdflower farm out small tasks to thousands of workers around the world. For a small fee (often a few cents) these workers will (for example) copy and paste website content. Another option is to collect simpler metrics at the source, such as counts of keywords, a common approach of many "big data" projects (Carneiro et al. 2009; Leskovec et al. 2009; Schmidt 2015).

2.2. From Text to Data

The content of each document must then be converted to quantitative data. Frequently, the objective is to create a term-document or term-frequency matrix where each row is a document and each column is a "feature" found in at least one of those documents.[2] Thus at this stage researchers need to decide on the appropriate unit of analysis. For example, presidential State of the Union addresses (SOUs) are lengthy and cover many different subjects. A project that examines SOU policy topics (e.g.) will probably be improved by parsing each address into more focused paragraphs or sentences.

The next step is to specify which features within each document will be used in the quantitative analysis. The starting point is usually to treat every unique word as a separate feature. Researchers then exclude document content that is thought to be irrelevant to the analysis and potentially misleading. Most text analysis packages include standard options that remove punctuation, common words (stopwords), very infrequent words (sparse terms), and word suffixes (stemming). However, each of these actions deserves careful consideration. For example, standard stopwords such as cant and cannot might be relevant features for a study of presidential address tone. The next step may be to create features beyond the basic bag of words. One common practice is to include word pairs (bigrams) as additional features. But the possibilities are truly endless. For example, instead of treating synonyms as separate words, a researcher might combine them into a single feature. They might also assign more weight to features that are thought to be especially informative, or create new features from outside information. Roberts et al. (2015) find that incorporating information about whether a blog has liberal or conservative leanings helps to predict blog topics.

2.3. Quantitative Analysis of Text

Simple metrics can be very useful and have the added virtues of transparency and replicability. Eggers and Spirling (forthcoming) study parliamentary dynamics by examining patterns in specific word usage across time. Casas, Davesa and Congosto (2016) assess

[2]The cell values indicate whether a feature is present (0,1) in a term-document matrix, or how often it is found (0,N) in a term frequency matrix.

media reporting of attitudes using lists of positive and negative words. However, much of the focus today (some would say hype) is on statistical machine learning methods. Scholars continue to debate (water cooler style) the differences between machine learning and statistics. We are certainly not going to settle that debate but we do think that the distinction can help to highlight general differences in approach. Political scientists are accustomed to using statistical methods to test theories. They choose the best model for the data (ordinary least squares, logistic regression, ...) before testing model specifications that include a limited number of theoretically-derived input (independent) variables. The focus is typically on the coefficients or parameters for the input variables - other things equal, are women significantly more likely to identify as Democrats than men? Whether the model accurately predicts the partisan identification of each voter is usually of secondary concern.

In machine learning research, the focus is usually on the outputs rather than the inputs. Instead of asking whether women are more likely to identify as Democrats, the objective is more likely to be to predict state level political opinion using Twitter (Beauchamp forthcoming). This focus on outputs leads researchers to be more concerned with prediction accuracy and less concerned with explanation. Beauchamp reports the features most associated with pro-Obama and pro-Romney poll shifts but does not try to explain why (for example) the most important predictive for Obama support is "75" and the most important for Romney is "cia." (Table 2). The focus on prediction also encourages more experimentation with different algorithms and features (Domingues 2015). We review some of the most relevant machine learning applications in section 3 of this article.

2.4. Evaluating Performance

Validation is a critical component of every text as data project (Saldana 2009; Grimmer and Stewart 2013). For some methods validation is straightforward. Supervised machine learning results are validated by comparing an algorithm's predictions to pre-existing "gold standard" results. These may be documents labeled by human annotators, but there are many other possibilities. The gold standard for Beauchamp are state level public opinion polls. In computer science, researchers frequently take advantage of on-line ratings and reviews to train and validate algorithms capturing sentiment. To guard against overfitting, researchers typically train the algorithm on one set of labeled examples before testing accuracy using a different, held-out, set.[3] Whether the gold standard labels do a good job of capturing the phenomenon of interest is a separate (and important) question. For other methods where no gold standard is available, validation can be more involved. For example, unsupervised machine learning methods (such as topic models) optimally cluster cases based on the latent structure of their features. Among other things, scholars have delved into specific examples within topics to show that the topics make sense; demonstrated that different algorithms produce similar clusters; and that variations in topic emphasis across time or venues correlate with real world events (Blei and Lafferty 2009; Quinn et al. 2010;

[3]Repeating this process several times, using different training and testing sets, and then aggregating the validation results (N-fold cross-validation), is an even better approach (Kohavi 1995; Arlot 2010).

Grimmer and King 2011; Roberts et al. 2014).

3. Recent Developments in Political Science

The purpose of this section is to provide a sense of the research opportunities available for political scientists. We make no attempt to be comprehensive but instead focus on four general research objectives. Two (classification and scaling) will be familiar to many readers (Grimmer and Stewart 2013). The other two (text reuse and semantics) have received less attention to date but represent important untapped research opportunities.

3.1. Classification

Classification is a popular objective of text as data projects. **Unsupervised machine learning** methods (e.g. K-means, Principal Components (PCA), Latent Dirichlet Allocation (LDA)) group documents based on co-occurring features. Despite their name, unsupervised methods require a lot of input from the user who must (among other things) specify the number of clusters in advance and interpret the results (by naming the clusters). In one of the earliest applications by political scientists, Quinn et al. (2010) used an unsupervised learner to classify Senate speeches for policy topic. They then validated their results by showing that their topics were similar to those developed using more time consuming methods. Bousaills and Coan (2016) and Farrell (2016) use topic modeling to investigate climate change "skepticism" in reports and communications by think tanks and interest groups. Grimmer and King (2011) demonstrate how unsupervised methods can lead to new discoveries. They find that congressional press releases cluster in ways that match Mayhews (1976) typology of constituent advertising, position-taking, and credit-claiming, but they also observe an additional cluster they label partisan taunting (see also Grimmer 2013). Roberts et al. (2014) show how incorporating additional information about documents (beyond the bag of words) into topics models can aid in interpretation of open-ended survey responses.

Whereas unsupervised methods are often used for discovery, **supervised learning** methods are primarily used as a labor saving device. For example, Workman (2015) and Collingwood and Wilkerson (2013) use supervised methods to apply a well established Policy Agendas topic coding system to new research domains (federal regulations and congressional bills). Boydstun et al. (2016) are currently labeling thousands of newspaper articles for issue frame with the long term goal of developing a supervised learner that can predict frames in other articles. The fact that supervised methods often require thousands of training examples makes them a non-starter for many researchers and projects. However, there are often creative ways to reduce the effort required. Examining 250,000 Enron emails, Drutman and Hopkins (2013) are able to use simple identification techniques to first exclude the 99% that were not political in nature. Crowdsourcing is also frequently used to build training sets in computer science. When a project does not require individual document labels, ReadMe is a supervised method that reliably predict class proportions using a much smaller number of training examples (Hopkins and King 2010). King et al. (2013) use ReadMe to classify millions of social media posts for topic in a study of government censorship in China. Ceron et al. (2013) use it to study citizens' policy

preferences in Italy and France.

Sentiment analysis is another important area of classification research where supervised and unsupervised methods are often used. The objective is to classify text ordinally (from negative to positive for example) rather than categorically. Because businesses care about how consumers are responding to their products on-line, sentiment analysis is a very hot and well funded area of research in computer science. As a result, political scientists can take advantage of many pre-existing training corpora for a wide variety of research domains.[4]

3.2. Scaling

Some of the earliest applications of automated text analysis in political science focused on using speeches and manifestos to locate European political parties in continuous ideological space (Laver et al. 2003; Slapin and Proksh 2008; Lowe 2008). Subsequent research has extended this research by employing new methods and investigating new domains. In a pathbreaking study, Benoit et al (2015) show that crowdsourcing can be a viable (preferred) alternative to expert-based approaches to locating parties on policy dimensions. Kluver (2009) uses statements by interest groups and EU regulators to estimate ideological positions and gauge influence. Diermeier et al. (2012) test several different approaches to estimating legislator ideology from statements in the Congressional Record (see also Lauderdale and Herzog 2015). Barbera (2015) uses Twitter data and information about their followers to estimate the ideological positions of politicians, parties, and individual citizens. Lauderdale and Clark (2015) combine past votes with topic modeling of judicial opinions to critique single dimensonal scaling of justices and to develop separate estimates of judicial ideology for different issue areas.

3.3. Text Reuse

Text reuse, as the name implies, is about discovering instances of similar language usage. The distinctive feature of text reuse algorithms is that they explicitly value word sequencing in judging document similarity. Political scientists have recently employed them to trace the origins of policy proposals in legislation (Wilkerson et al. 2015), to study the influence of interest groups in state legislatures (Hertel-Fernandez and Kashin 2015),[5] and to study party messaging strategies (Jansa et al. 2015). Other possibilities yet to be exploited by political scientists include studying the diffusion of political memes and contagion effects in new and old media (Leskovec et al. 2009; Smith et al. 2013). Different algorithms also support different types of analyses. Global alignment algorithms (e.g. Needleman-Wunsch 1970) measure the overall similarity of documents whereas local alignment algorithms (e.g. Smith Waterman 1981) identify and score shared word sequences within documents. Thus, in a study of lawmaking or treaty negotiations, a global alignment algorithm might be used to compare how much the entire proposal changes as it moves from one stage of the process to the next, whereas a local alignment algorithm could be used to investigate the

[4]For example:
`http://www.cs.cornell.edu/home/llee/data/`
`http://mpqa.cs.pitt.edu/corpora/mpqa_corpus/`

[5]see also the Legislative Influence Detector project https://dssg.uchicago.edu/lid/

fates of more specific provisions or proposals.

3.4. Natural Language Processing (NLP)

Social network analysis often employs text to investigate relationships among actors (Ward et al. 2011). Natural language processing makes it possible to move beyond simply establishing connections to investigating the state of relationships, for example by moving from "whom" to "who did what to whom? (Van Atteveldt et al, 2016). For example, political event data analysis draws on media reports to systematically monitor interactions between international actors. Instead of simply counting the number of times two actors are mentioned in reports, event data analysis incorporates syntax (sentence structure) and semantics (word meaning) to systematically track whether a relationship is improving or worsening and (possibly) to attribute credit or blame for developments.

Early event data research relied on human annotators to develop dictionaries of named entities and actions (Schrodt and Gerner 1994; Gerner and Schrodt 1994; Schrodt et al. 1994). More recent research seeks to dramatically expand the scope of this research by taking advantage of extensive NLP resources developed by computer scientists and linguists (Leetaru and Schrodt 2013; see Ward et al. 2013 for an overview). For example, the Stanford Parser and the Stanford Named Entity Recognizer can be used to automatically extract specific parts of speech from documents and to tag different references the same entity (USA, America, United States etc.). Other valuable resources such as Wordnet can be used to identify synonyms for similar actions or sentiment. Sagin et al. (2013) show how semantic information can be used to reduce many different ways of talking about terrorism in debates to a more limited set of issue frames. Denny et al. (2015) demonstrate how NLP methods can be used to systematically isolate the substantive provisions in legislation that typically includes lots of irrelevant "boilerplate" language. The creative possibilities are extensive and Bird et al (2009) provide an excellent primer on available NLP resources.

4. Topic Model Instability and a Call for Greater Attention to Robustness in Text as Data Research

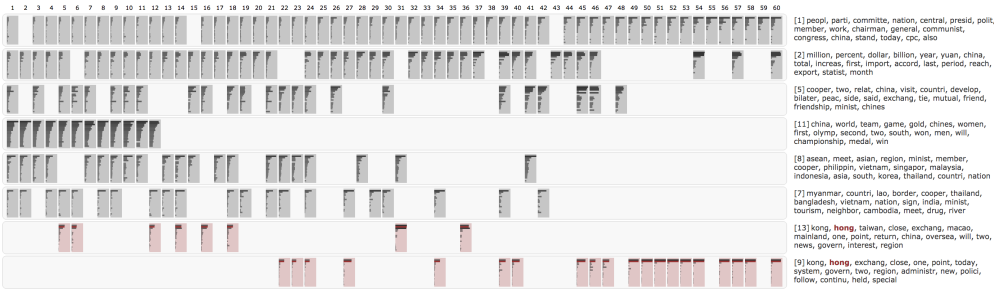
In this final section, we shift from providing an overview of the field to delving into one contemporary challenge in more detail. Unsupervised machine learning methods (topic models) are very popular in political science in part because they classify documents without the extensive labeling efforts often required for supervised learning methods. The common practice has been to report and validate a single topic model after comparing results for several different models that vary by the number of topics specified by the researcher. This choice is usually based on the researcher's subjective judgment about which model's clusters best reflects the substantive goals of the project.[6]

[6]Supplemental and replication materials for this section can be found on the Supplemental Materials section of the Annual Review website and at `https://github.com/CasAndreu/wilkerson_casas_2016_TAD`. These materials include a `python` module, `rlda`, to apply the Robust Latent Dirichlet Allocation models used here (`https://github.com/CasAndreu/rlda`)

Earlier we noted that the absence of a gold standard makes validation more challenging for these methods. However, a second challenge is model instability. Chuang et al. (2015) illustrate this challenge by estimating the `same` structural topic model 50 times to find that only two of 25 topics persist across all of the estimations (Figure 1 on page 179). This can happen because different estimations can converge at different local maxima (Roberts et al. 2015). In a second experiment, the same authors find that manipulating just one feature of a structural topic model also leads to very different results. Many machine learning packages remove rarely used words by default to reduce processing time and avoid overfitting. In Figure 1, each of the 50 columns is a 25 topic LDA model where the only difference is the threshold used to exclude sparse terms. Each row is a topic. The shaded cells indicate when a model includes the topic.[7] What is remarkable is the lack of consensus in terms of how the cases should be clustered when only one feature that is generally assumed to be unimportant is varied.

Figure 1

Impact of a feature on topic stability (from Chuang et al. 2015)



A number of recent studies have proposed different ways to assess and respond to topic model instability (Grimmer and King 2011; Schmidt 2013; Graber et al. 2014; Roberts et al. 2014). However, the focus, as far as we are aware, continues to be on selecting and validating a single best model. We think that recent ?ndings concerning model instability will make it increasingly di?cult to publish results of this kind. In conventional statistical studies, researchers try to demonstrate that their results are robust by reporting results for multiple model specifications. A study examining gender and voting will test and report several combinations of theoretically-derived independent variables. Supervised machine learning analyses also commonly address robustness by basing results on the consensus prediction of an ensemble rather than a single algorithm. Grimmer and King (2010) propose a method for comparing topic model results for different algorithms, but they do not incorpoates those differences into their findings.

Robustness can be evaluated with respect to methods, parameters, features, and data partitions. No study can consider all permutations but we do think that political scientists using text as data methods should explicitly address robustness in their results: do the

[7] They assume that two models share the same topic if the cosine similarity of the topic terms is greater than .9. Darker shades in the figure indicate higher similarity.

central findings stand up to reasonable variations in modeling choices? Where topic models are concerned, one option is to move away from the current convention of reporting results for a single model. Grimmer and King (2010) propose a method for comparing topic model results for different algorithms, but they do not incorporate those differences into their findings.

4.1. Exploring the topics of legislators' floor speeches

In this section we illustrate how topic robustness can inform a study of congressional floor speeches. Members of the U.S. House of Representatives gave almost 10,000 "one-minute" floor speeches during the 113th Congress (2013-14). These speeches are given before ordinary business and are primarily intended for public consumption (Schneider 2015)[8]. A quick review indicates that their subjects are often quite diverse. Some honor constituent accomplishments (such as a state basketball championship), whereas others address political and legislative issues. However, to our knowledge, no one has systematically investigated what members talk about in these speeches. What topics are covered and which are the most common? Do Republicans and Democrats tend to talk about the same issues or emphasize different ones?.

To examine these questions we first used the Sunlight Foundation's *Capitol Words API* to download all member statements from the *Congressional Record* of the 113th Congress. We then removed statements that did not begin with the opening phrase of a one minute speech: "Mr. Speaker, I rise today..." This produced a corpus of 5,346 one minute speeches given by 179 Democrats and 4,358 given by 213 Republicans. We converted the words in each speech to lower case, removed punctuation, stopwords, word stems, and words of two characters or less. Finally we constructed a term document matrix where each row is a one-minute speech and each column is a vector indicating whether a feature/word is present in a given speech.

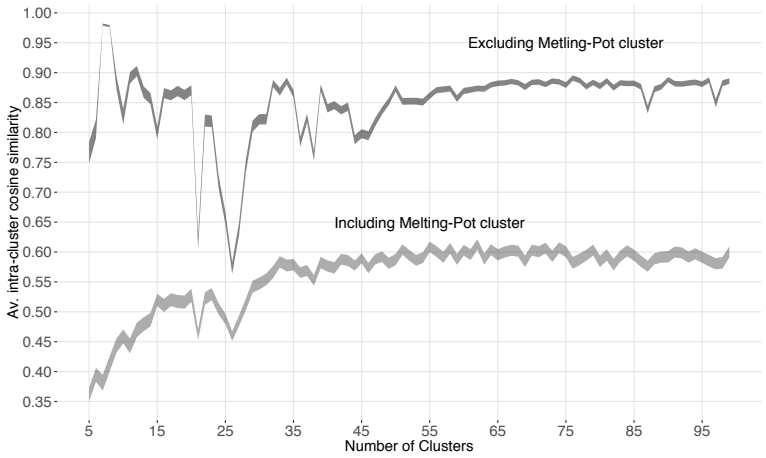
The next step was to estimate a series of Latent Dirichlet Allocation (LDA) models where the number of topics (k) ranges from 10 to 90 in 5 topic increments (Blei et al. 2003). These 17 models yield 850 topics (10 + 15 + 20... + 90). To determine which topics were robust we first calculated cosine similarity[9] for all topic pairs (resulting in 722,500 similarity scores) and then used the Spectral Clustering algorithm to group the 850 topics based on cosine similarity. The Spectral Clustering algorithm does this by maximizing average intra-cluster cosine similarity for a given number of clusters c. The substance of a given cluster can then be investigated by examining the most predictive words ("top terms") in each cluster.

[8] https://www.fas.org/sgp/crs/misc/RL30135.pdf

[9] For each possible pair of topics, $cos(\theta) = \frac{a \cdot b}{\|a\| \, \|b\|}$, where a and b are vectors of counts recording topic-word assignments in the final estimation iteration.

igure 2

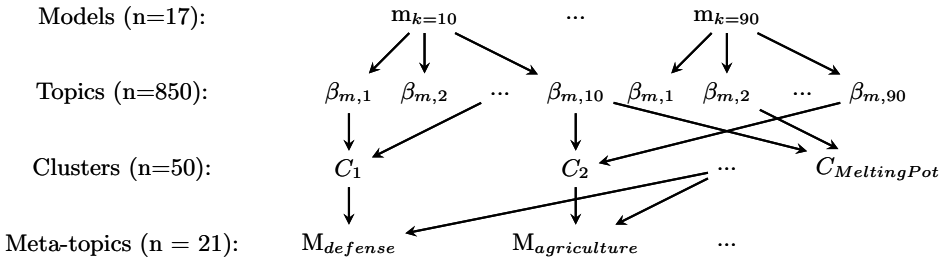
Number of clusters and average intra-cluster similarity



Ten thousand speeches by 435 lawmakers should cover a diverse set of topics. On the other hand, dividing the speeches into too many clusters may complicate the analysis without improve the overall fit (average intra-cluster similarity) of the model to the data. Figure 2 displays how fit improves as the number of speech clusters c is varied from 5 to 100. The lower line indicates that more clusters improve overall fit until at about 50 clusters. The upper line indicates that the average similarity of the clusters excluding the largest catch all "melting pot" cluster is also quite volatile until about 50 clusters. We therefore base our analysis on the robust topics from a 50 cluster model.

Figure 3

Workflow of moving from 17 topic models to 21 meta-topics

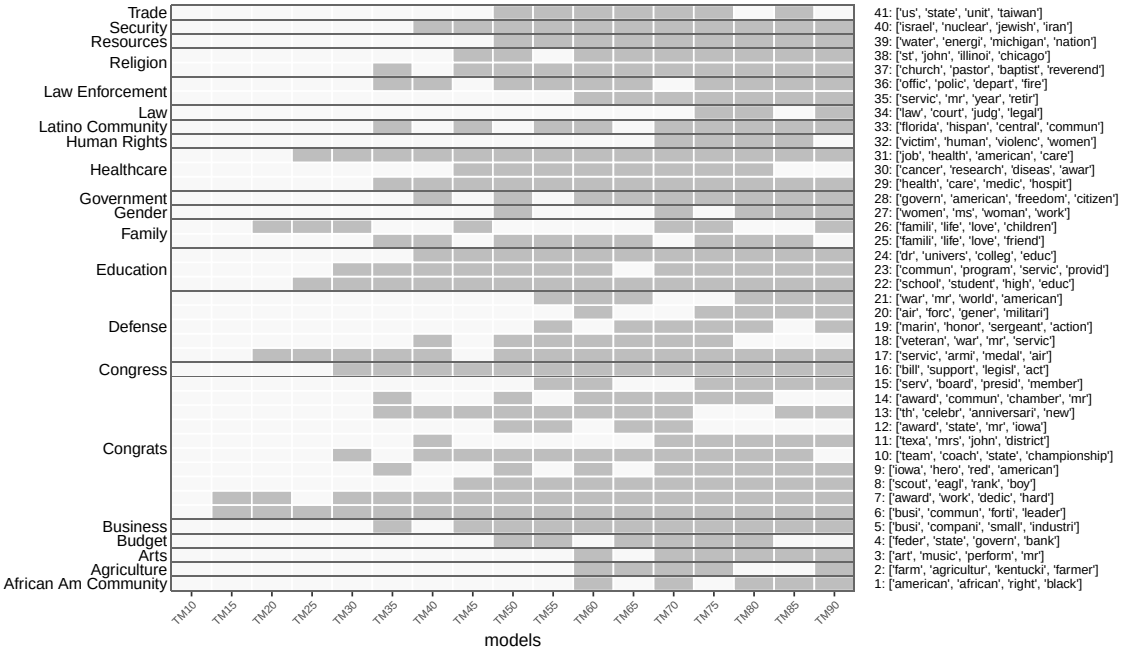


After clustering the 850 topics from 17 models into 50 clusters (see the above workflow diagram), we then grouped some of the clusters into "meta-topics." For example, the "education" metatopic includes three clusters about education. We also exclude all of the topics within one "unclear" meta-topic where we were unable to discern a consistent theme. Thus the results presented below are based on 697 of the original 850 topics from 16 of the 17 original topic models. In figure 4 each column is one of the 17 LDA topic models (ranging from 10 to 90 topics) and each row is a topic cluster. The 21 substantive

meta-topics are listed on the left. The shaded cells indicate where the topics in each cluster or meta-topic originate. The education meta-topic, for example, includes 37 topics found in 14 different topic models. In our view the figure underscores the drawbacks of presenting results based on a single model. Topics that are common to many models are often missing from any one of them.

Figure 4

The 21 metatopics of a 50 cluster model

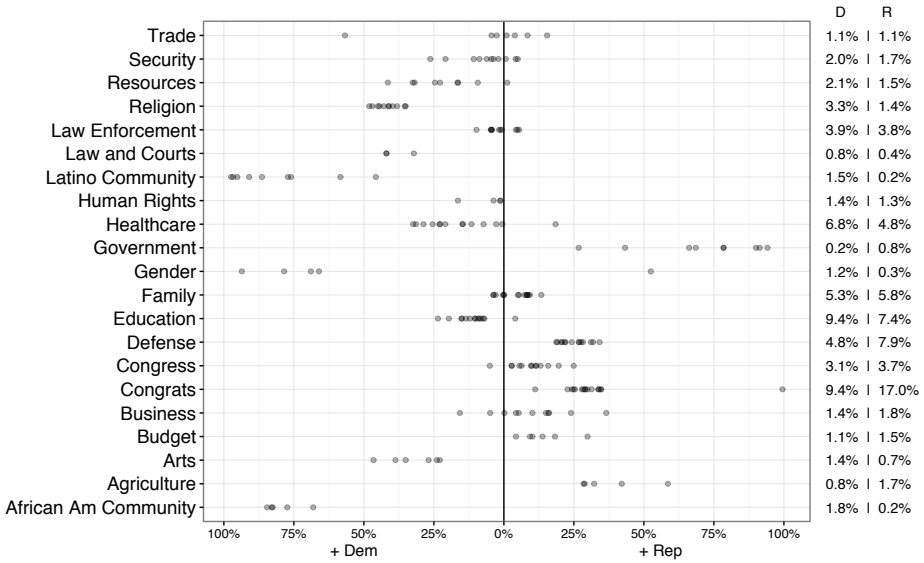


4.2. Topic attention in one-minute speeches

An LDA topic represented by a vector expressing the probability of each unique word belonging to that topic (see Blei et al. 2003, pp.996-997). Each document is assumed to be about each topic with some positive probability. To study speech attention we must first label individual speeches for primary topic. We assume each speech is about its most probable topic. Thus we classify 9,704 speeches for each of 16 topic models. We then report results for just those topics from each model that are part of the 21 meta-topics.

igure 5

A robust examination of issue emphasis in one-minute speeches



Each row in figure 5 is one of the 21 metatopics. Each dot is a result for one topic model. For example, for education, the consensus of the different topic models is that Democrats gave more speeches about education than Republicans. It is reassuring that the models generally agree concerning partisan emphasis for most of the meta-topics. On the other hand, there is often considerable disagreement regarding the size of the difference.

The average amount of attention (across the models) given to different metatopics (such as education) by Republicans and Democrats are displayed on the far right. The results indicate that Republicans were most likely to give Congratulatory speeches (17%), followed by speeches about Defense (8%), Education (7%) and Family (6%). Democrats also gave lots of Congratulatory speeches (9%), but were as likely to give speeches about Education (9%), followed by Healthcare (7%) and Family (5%).

4.3. Validation

We think that similar estimates of topic emphasis across different models is an important type of validation. Figure 5 should inspire confidence in terms of general differences in speech topic emphasis, but less confidence in the amount of difference in many cases. It is also worth noting that our results generally support Petrocik's (1996) "issue ownership" argument. The main exception seems to be agriculture. According to Petrocik, Democrats "own" agriculture whereas most of the models of our analysis indicate that Republicans own it. We therefore took a closer look at who was giving speeches about agriculture and found a strong correlation between the proportion of a member's speeches that were about agriculture and the number of district workers employed in the Agriculture, Forestry, Fishing,

Hunting and Mining industries (Pearson's $r = 0.4$).[10] Thus it seems likely that there has been a transfer of ownership on this issue since Petrocik's article was published 20 years ago.

5. Discussion

Computerized text analysis is transforming political science research because scholars now have the ability to explore massive amounts of politically relevant text using increasingly sophisticated tools. These developments have already produced important advances in research methods (Hopkins and King 2010; Benoit et al. 2015), opened the door to new research questions (Wilkerson et al. 2015), and altered current understandings (Lauderdale et al. 2015). We have argued that researchers do not need to be computer programmers or statistical methodologists to use text as data methods in their research. They do need to be attentive to the same concerns about validity and reliability that apply to all methods.

Unsupervised learning methods (such as topic models) have proven to be one of the more popular methods used in political science. We believe that a central attraction for many researchers is that they do not require labeled training sets. This appeal is understandable and supervised learning methods have their own limitations. However, the absence of any gold standard makes choosing and validating a model even more challenging for unsupervised methods.

To date, unsupervised learning research has largely focused on better ways to validate a single topic model chosen by the researcher after first comparing results for several different models. We think that recent findings concerning model instability will make it increasingly difficult to publish results of this kind. In the last part of this paper, we propose a robust approach to reporting topic model results that takes advantage of the information provided by alternative specifications. This approach has its own limits, but in our view it is informative and transparent and adheres to current conventions that lead researchers to explictly address robustness in statistical studies.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors thank Jeffrey Arnold, Noah Smith, Jason Chuang, and an anonymous reviewer for helfpul feedback. They are of course not responsible for any errors or oversights in this article.

[10] see the on-line appendix for more details

LITERATURE CITED

. Alvarez, R. Michael, ed. 2016. *Computational Social Science: Discovery and Prediction. Analytical Methods for Social Research.* New York, NY: Cambridge University Press.

. Barbera, Pablo. 2015. Birds of the Same Feather Tweet Together. Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis*, 23(1):76−91.

. Benoit, Kenneth, Drew Conway, Benjamin Lauderdale, Michael Laver, and Slava Mikhaylov. 2015. Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data. *American Political Science Review, forthcoming.*

. Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python - Analyzing Text with the Natural Language Toolkit.* OReilly.

. Bishop, R. L., and M. L. Weisfeldt. 1976. Sodium Bicarbonate Administration during Cardiac Arrest. Effect on Arterial pH PCO2, and Osmolality. *JAMA* 235 (5): 506−9.

. Boydstun, Amber, Ross Butters, Dallas Card, Justin Gross, Philip Resnik, and Noah Smith. 2016. Under What Conditions Does Media Framing Influence Public Opinion on Immigration? Paper presented at the *Midwest Political Science Association Annual Meeting* in Chicago.

. Breiman, Leo. 2001. Random Forests. *Machine Learning* 45 (1): 5−32.

. Cardie, Claire, and John Wilkerson. 2008. Text Annotation for Political Science Research. *Journal of Information Technology & Politics* 5 (1): 1−6.

. Carneiro, Herman Anthony, and Eleftherios Mylonakis. 2009. Google Trends: A WebBased Tool for RealTime Surveillance of Disease Outbreaks. *Clinical Infectious Diseases* 49 (10): 1557−64.

. Casas, Andreu, Ferran Davesa, and Mariluz Congosto. 2016. The Media Coverage of a Connective Action: The Interaction between the 15-M Movement and the Mass Media. *Revista Espanola de Investigaciones Sociologicas* 155:73−96.

. Ceron, A., L. Curini, S. M. Iacus, and G. Porro. 2014. Every Tweet Counts? How Sentiment Analysis of Social Media Can Improve Our Knowledge of Citizens Political Preferences with an Application to Italy and France. *New Media & Society* 16 (2): 340−58.

. Chang, Jonathan, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *Proceedings of the Neural Information Processing Systems Conference.*

. Chuang, Jason, Margaret Roberts, Brandon Stewart, Rebecca Weiss, Dustin Tingley, Justin Grimmer, and Jeff Heer. 2015. TopicCheck: Interactive Alignment for Assessing Topic Model Stability. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, 175−84. Denver, Colorado: Association for Computational Linguistics.

. Chuang, Jason, John D. Wilkerson, Rebecca Weiss, Dustin Tingley, Brandon M. Stewart, Margaret E. Roberts, Forough Poursabzi-Sangdeh, et al. 2014. "Computer-Assisted Content Analysis: Topic Models for Exploring Multiple Subjective Interpretations". Proceedings of *Advances in Neural Information Processing Systems Workshop on Human-Propelled Machine Learning.*,Montreal, Canada.

. Collingwood, Loren, and John Wilkerson. 2011. Tradeoffs in Accuracy and Efficiency in Supervised Learning Methods. *Journal of Information Technology & Politics*, 4: 1−28.

. Dice, Lee R. 1945. Measures of the Amount of Ecologic Association Between Species. *Ecology* 26 (3): 297−302.

. Diermeier, D., B. Yu, S. Kaufmann, and J.E. Godbout. 2012. Language and Ideology in Congress. *British Journal of Political Science* 42 (1):31−55.

. Domingos, Pedro. 2015. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. New York: Basic Books, a member of the Perseus Books Group.

. Drutman, Lee, and Daniel J. Hopkins. 2013. The Inside View: Using the Enron Email Archive to Understand Corporate Political Attention. *Legislative Studies Quarterly* 38 (1):5−30.

. Eggers, Andrew, and Arthur Spirling. 2016. The Shadow Cabinet in Westminster Systems: Modeling Opposition Agenda Setting in the House of Commons, 1832-1915. *British Journal of Political Science, forthcoming*.

. Fenno, Richard F. 1978. *Home Style: House Members in Their Districts*. New York: HarperCollins.

. Gerner, Deborah J., Philip A. Schrodt, Ronald A. Francisco, and Judith L. Weddle. 1994. Machine Coding of Event Data Using Regional and International Sources. *International Studies Quarterly* 38 (1): 91.

. Grimmer, Justin, and G. King. 2011. General Purpose Computer-Assisted Clustering and Conceptualization. *Proceedings of the National Academy of Sciences* 108 (7): 2643−50.

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

. Grimmer, Justin. 2013. Appropiators Not Position Takers: The Distorting Effects of Eletoral Incentives on Congressional Representation. *American Journal of Political Science* 57 (3): 624−42.

. Herterl-Fernandez, Alexander, and Konstantine Kashin. 2015. Capturing Business Power across the States with Text Reuse. Paper presented at the *Midwest Political Science Association Annual Meeting* in Chicago.

. 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.

. Huang, A. 2008. Similarity Measures for Text Document Clustering. In *Proceedings of the Sixth New Zeland Computer Science Research Student Conference*, 49−56. Christchurch, New Zeland.

. Jansa, Joshua, Eric Hansen, and Virginia Gray. Copy and Paste Lawmaking: The Diffusion of Policy Language across American State Legislatures. *Working Paper*.

. Jockers, Matthew Lee. 2014. *Text Analysis with R for Students of Literature*. Cham: Springer.

. King, Gary, Jennifer Pan, and Margaret E. Roberts. 2013. How Censorship in China Allows Government Criticism but Silences Collective Expression. *American Political Science Review* 107 (02): 326−43.

. Kluver, H. 2009. Measuring Interest Group Influence Using Quantitative Text Analysis. *European Union Politics*10 (4): 535−49.

. Klver, Heike. 2013. Lobbying as a Collective Enterprise: Winners and Losers of Policy Formulation in the European Union. *Journal of European Public Policy* 20 (1): 59−76.

. Kohavi, Ron. 1995. A Study of Cross-Validation and Bootsrap for Accuracy Estimation and Model Selection. In *Proceedings of the International Join Conference on Artificial Inteligence*.

. Lauderdale, Benjamin, and Alexander Herzog. *Forthcoming*. Scaling Politically Meaningful Dimensions Using Texts and Votes. *Political Analysis*.

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

. Leetaru, Kalev, and Philip Schrodt. 2013. Gdelt: Global Data on Events, Location, and Tone, 19792012. Paper presented at the *ISA Annual Convention*.

. Leskovec, Jure, Lars Backstrom, and Jon Kleinberg. 2009. Memetracking and the Dynamics of the News Cycle. In *KKD* 497−506. Paris, France.

. Lo, James, Sven-Oliver Proksch, and Jonathan B. Slapin. 2014. Ideological Clarity in Multi-Party Competition: A New Measure and Text Using Election Manifestos. *British Journal of Political Science*. 46(3):591−610.

. Lowe, W. 2008. Understanding Wordscores. *Political Analysis* 16 (4): 35671.

. Mayhew, David R. 1974. Congress: The Electoral Connection. New Haven: Yale University Press.

. Monroe, B. L., and P. A. Schrodt. 2008. Introduction to the Special Issue: The Statistical Analysis of Political Text. *Political Analysis* 16 (4): 35155.

. Needleman, Saul B., and Christian D. Wunsch. 1970. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal of Molecular Biology* 48 (3): 44353.

. Pang, Bo, and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. Now Publishers Inc.

. Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume* 10, 7986.

. Petrocik, John R. 1996. Issue Ownership in Presidential Elections, with a 1980 Case Study. *American Journal of Political Science* 40 (3): 82550.

. Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science* 54 (1): 20928.

. 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: Structural Topic Models for Survey Responses. *American Journal of Political Science* 58 (4): 106482.

. Roberts, Margaret, Brandon Stewart, and Dustin Tingley. 2016. Navigating the Local Modes of Big Data: The Case of Topic Models. In *Computational Social Sciences*, edited by R. Michael Alvarez. New York, NY: Cambridge University Press.

. Saldana, Johnny. 2009. *The Coding Manual for Qualitative Researchers*. Los Angeles, Calif: Sage.

. Schmidt, Benjamin. 2013. Words Alone: Dismantling Topic Models in the Humanities. Journal of Digital Humanities. (2)1.

. Schmidt, Benjamin. 2015. Is It Fair to Rate Professors Online? *The New York Times*, December 16, sec. Room for debate.

. Schrodt, Philip A., and Deborah J. Gerner. 1994. Validity Assessment of a Machine-Coded Event Data Set for the Middle East, 1982-92. *American Journal of Political Science* 38 (3): 825.

. Slapin, Jonathan B., and Sven-Oliver Proksch. 2008. A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science* 52 (3): 70522.

. Smith, David A., Ryan Cordell, and Elizabeth Maddock Dillon. 2013. Infectious Texts: Modeling Text Reuse in Nineteenth-Century Newspapers. In *Proceedings of the IEEE Conference*, 8694.

. Smith, T.F., and M.S. Waterman. 1981. Identification of Common Molecular Subsequences. *Journal of Molecular Biology* 147 (1): 19597.

. Wallach, Hanna, Lee Dicker, and Shane Jensen. 2010. An Alternative Prior for Nonparametric Bayesian Clustering. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol. 9. 89299.

. Ward, Michael, Andreas Beger, Cutler Josh, Matthew Dickenson, Cassy Dorff, and Ben Radford. 2013. "Comparing GDELT and ICEWS Event Data. *Analysis* 21: 26797.

. Ward, Michael, Katherine Stovel, and Audrey Sacks. 2011. "Network Analysis and Political Science," *Annual Review of Political Science,* 14:245−264

. Wilkerson, John, David Smith, and Nicholas Stramp. 2015. Tracing the Flow of Policy Ideas in Legislatures: A Text Reuse Approach" *American Journal of Political Science*, 59(4):943−956.

. Yogatama, Dani, and Noah Smith. 2015. Bayesian Optimization of Text Representations. In arXiv:1503.00693.