Introduction to the Virtual Issue: Recent Innovations in Text Analysis for Social Science

Margaret E. Roberts*

1 Text Analysis for Social Science

In 2008, *Political Analysis* published a groundbreaking special issue on the analysis of political text, examining some of the initial efforts in political science to consider text as a data source and to develop methods for analyzing text data.[1] In their introduction to the special issue, Monroe and Schrodt (2008) note that text – one of the most common mediums through which political phenomenon are documented – is underutilized in the social sciences and they argue for further research. They suggest the research discussed in the special issue should be a jumping-off point, or "departure lounge" for future text as data research.

Answering their call, in the last eight years, the field of "text as data" in social science has grown dramatically. As the number of sources and types of textual data documenting social science phenomenon has exploded, so too have methods for, and the use of, text analysis in social science research. The articles included in this virtual issue of *Political Analysis* showcase how the study of text analysis in political science has built on these initial political science approaches. This virtual issue includes:

1. Grimmer, Justin, and Brandon M. Stewart. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political Analysis* 21.3 (2013): 267-297.

2. D'Orazio, Vito, Steven T. Landis, Glenn Palmer, and Philip Schrodt."Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines." *Political Analysis* 22.2 (2014): 224-242.

3. Lowe, Will, and Kenneth Benoit. "Validating Estimates of Latent Traits from Textual Data Using Human Judgment as a Benchmark." *Political Analysis* 21.3 (2013): 298-313.

4. Grimmer, Justin. "A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases." *Political Analysis* 18.1 (2010): 1-35.

*Assistant Professor, Department of Political Science, University of California, San Diego, Social Sciences Building 301, 9500 Gilman Drive, #0521, La Jolla, CA 92093, meroberts@ucsd.edu, MargaretRoberts.net

[1] The articles in the special issue included Monroe and Schrodt (2008); Lowe (2008); Monroe, Colaresi and Quinn (2008); Bailey and Schonhardt-Bailey (2008); Klebanov, Diermeier and Beigman (2008); Van Atteveldt, Kleinnijenhuis and Ruigrok (2008)

5. Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley. "Computer-Assisted Text Analysis for Comparative Politics." *Political Analysis* 23.2 (2015): 254-277.

6. Elff, Martin. "A dynamic state-space model of coded political texts." *Political Analysis* 21.2 (2013): 217-232.

7. Harris, J. Andrew. "What's in a Name? A Method for Extracting Information about Ethnicity from Names." *Political Analysis* 23.2 (2015): 212-224.

The authors in this virtual issue have enhanced the tools for text as data by providing methods that allow for the analysis of more types of text data and identifying new points in the research process where text analysis can be used. The papers included in this virtual issue have developed frameworks for the use of textual data (Grimmer and Stewart, 2013), developed methods for document sampling (D'Orazio et al., 2014) to validation (Lowe and Benoit, 2013), enhanced the use of non-textual metadata in text analysis (Grimmer, 2010; Lucas et al., 2015), and improved upon existing approaches to allow textual data to travel across time, countries and languages (Lucas et al., 2015; Elff, 2013; Harris, 2015). These new methods have broadened the scope of text methods in political science and expanded their accessibility across subfields of political science. Many of these papers are accompanied by extensively documented software, making them easy to use for applied researchers.

2 A Framework and Principles for the Analysis of Text

The first paper in this virtual issue provides a framework and template for understanding text analysis in the social science research process. Grimmer and Stewart (2013) is a must-read for any social scientist interested in using text analysis in their research. In a flow-chart of text analysis methods, Grimmer and Stewart provide a map of the text analysis toolkit, from the acquisition and preprocessing of text, to general approaches for estimating known categories of interest from text, to more exploratory approaches for researchers wishing to describe the contours of their data. Laying out four principles of automated text analysis, they caution readers to be wary of the pitfalls of automated text analysis: importantly that text analysis methods are not meant to replace, but rather to augment humans – reading is still necessary! – and that extensive validation of methods for text analysis is necessary to ensure that researchers are not misled by models that are necessarily much simpler than the text they analyze.

3 Sampling and Validation

The next two papers in the issue create methods for text analysis at essential points in the research process that are often overlooked by applied researchers. D'Orazio et al. (2014) take up the question of text retrieval: how can a researcher with a large amount of text data sort through the data to extract the text she is interested in studying? The authors propose a two-stage support vector machine (SVM) workflow where documents defined by a broad search are first coded into relevant and not relevant sets, and then SVM is used to distinguish relevant documents from those that are irrelevant. The authors apply this innovative approach to the Militarized Interstate Dispute (MID) dataset by using their method to find incidences of conflict among vast numbers of news reports. This method improves the efficiency and accuracy of finding relevant documents with respect to the previously used method of human coding and thus decreases bias in any subsequent analysis of the dispute dataset.

In another innovative paper moving text to another place in the research process, Lowe and Benoit (2013) develop a validation procedure to verify that ideological scalings of text reflect human perceptions of these ideologies. Responding to Grimmer and Stewart's (2013) call for validation of text models, the authors suggest a method where human coders evaluate pairs of documents and then can be scaled in order to compare the output of the human coding to the estimates produced by an ideological scaling model. The authors apply this method to legislative debates about the 2010 Irish budget, and compare human evaluations of pro- versus anti-budget speeches to ideological scores produced by the algorithm Wordfish (Proksch and Slapin, 2010). The algorithmic and human measures reassuringly largely correlate, except in the instance of one party. This deviation between the algorithm and humans allows the authors to identify the ways in which the text model is useful and the ways in which it fails to capture the nuances of the text.

4 Incorporating Metadata into Models of Text

The following two papers in the issue expand the types of data that can be used in conjunction with text analysis. While most unsupervised methods of text rely on simply the words within the text to sort and bin the data, these methods allow for the inclusion of detailed metadata associated with documents, such as information about the author, time period, or publication.

Grimmer (2010) introduces the Expressed Agenda Model, a Bayesian hierarchical topic model designed to estimate the topical content of statements made by political actors. This single-membership topic model acknowledges that the topical content of text are naturally sorted by author – authors will be more likely to discuss the topics they have before – but that topics themselves are general across senators. Thus the model incorporates the information about the texts' authors, estimating the topics each author is likely to focus on across texts. Grimmer (2010) applies the model to estimate the political priorities of members of Congress using 24,000 Senate press releases, providing one of the first comprehensive analyses of the topics that senators are most likely to focus on in statements to their constituents.

Building of the insights in Grimmer (2010) for incorporating document metadata with topic models, Lucas et al. (2015) provide a framework for topic models in comparative politics. In particular, they focus on the Structural Topic Model (STM) (Roberts et al., 2014), which builds off of Grimmer (2010) by allowing for the inclusion of arbitrary document-level covariates in a mixed-membership topic model. STM estimates the relationship between these covariates and topical prevalence, or the amount the document discusses a topic, and topical content, or the way in which a document discusses a topic. Including covariates allows topics to be estimated at the corpus-level, while providing flexibility for deviations in the amount and way in which a topic is discussed by covariate information such as author, time, or political party.

5 Enabling Text Analysis to Travel

Lucas et al. (2015) show how the incorporation of metadata in STM allow for the estimation of topic models in multilingual corpuses. Fist, they translate a multilingual corpus of text into a common language by machine translation tools. Using STM, they include the metadata on the document's original language in the topic model estimation to account for machine translation errors. They use this approach to analyze Chinese and Arabic microblog data to understand how social media users around the world reacted to the Edward Snowden revelations.

Also allowing the analysis of text to travel between countries, Elff (2013) develops a dynamic state-space model to estimate political positions of parties from text, applying the new method to statements of electoral positions in countries in the West, compiled and coded by the Comparative Manifestos Project (CMP) (`https://manifestoproject.wzb.eu/`). While the CMP estimates the positions of political parties by the amount of time each party spends discussing topics within the text, Elff (2013) observes that the amount of time a party spends discussing an issue may be related both to the party's position on the issue, but also to the salience of the issue during the time period. In a particularly innovative twist, Elff (2013) also allows for the positions of parties to move over time, and Elff (2013) provides estimates of the evolution of party positions across multiple countries and time periods.

One of the difficulties of studies in comparative politics is the lack of reliable data on even the most basic demographics, like ethnicity. Harris (2015) provides a method for estimating the ethnicities of names when data about ethnicities are not available. Following King and Lu (2008) and Hopkins and King (2010), Harris (2015) focuses not on estimating the ethnicity for each individual name, but rather on estimating the proportion of people from various ethnic groups for a set of names. Validating the estimates in North Carolina where ground-truth data is known, Harris (2015) applies the method to estimate ethnic displacement using the names in voter registration records in Kenya, where the data on ethnic composition is not available.

6 Concluding Remarks

Over the last eight years, political scientists have pushed the envelope of text analysis methods as applied to social science, providing a general framework for the applied researcher, expanding the use of text analysis to different points in the research process, allowing for the inclusion of metadata, and pushing text analysis to travel to across languages and countries. This virtual issue provides a sampling of the innovations in text as data in political science and clues as to where the field is going. First, the integration of text with outside metadata highlighted in these papers suggests a potential for further integration between types of data in political science research. Many of the same methods for integrating text and traditional political science datasets could be generalized to other types of high-dimensional data, like images or audio. Second, text could be used in still other areas of the research process, outside of sampling and measurement. Further research needs to be done to explore the ability of text to test causal effects and to optimize text research for qualitative exploration and discovery.

7 About the Author

Margaret Roberts is an Assistant Professor of Political Science at University of California, San Diego. She has worked on a variety of methods and applications for automated content analysis. Her papers are available at `http://margaretroberts.net`.

References

Bailey, Andrew and Cheryl Schonhardt-Bailey. 2008. "Does deliberation matter in FOMC monetary policymaking? The Volcker Revolution of 1979." *Political Analysis* 16(4):404–427.

D'Orazio, Vito, Steven T Landis, Glenn Palmer and Philip Schrodt. 2014. "Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines." *Political Analysis* 22(2):224–242.

Elff, Martin. 2013. "A dynamic state-space model of coded political texts." *Political Analysis* 21(2):217–232.

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):267–297.

Harris, J Andrew. 2015. "What's in a Name? A Method for Extracting Information about Ethnicity from Names." *Political Analysis* 23(2):212–224.

Hopkins, Daniel J and Gary King. 2010. "A method of automated nonparametric content analysis for social science." *American Journal of Political Science* 54(1):229–247.

King, Gary and Ying Lu. 2008. "Verbal autopsy methods with multiple causes of death." *Statistical Science* 23(1):78–91.

Klebanov, Beata Beigman, Daniel Diermeier and Eyal Beigman. 2008. "Lexical cohesion analysis of political speech." *Political Analysis* 16(4):447–463.

Lowe, Will. 2008. "Understanding wordscores." *Political Analysis* 16(4):356–371.

Lowe, Will and Kenneth Benoit. 2013. "Validating Estimates of Latent Traits from Textual Data Using Human Judgment as a Benchmark." *Political Analysis* 21(3):298–313.

Lucas, Christopher, Richard A Nielsen, Margaret E Roberts, Brandon M Stewart, Alex Storer and Dustin Tingley. 2015. "Computer-Assisted Text Analysis for Comparative Politics." *Political Analysis* 23(2):254–277.

Monroe, Burt L, Michael P Colaresi and Kevin M Quinn. 2008. "Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict." *Political Analysis* 16(4):372–403.

Monroe, Burt L and Philip A Schrodt. 2008. "Introduction to the Special Issue: The Statistical Analysis of Political Text." *Political Analysis* 16(4):351–355.

Proksch, Sven-Oliver and Jonathan B Slapin. 2010. "Position taking in European Parliament speeches." *British Journal of Political Science* 40(03):587–611.

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):1064–1082.

Van Atteveldt, Wouter, Jan Kleinnijenhuis and Nel Ruigrok. 2008. "Parsing, semantic networks, and political authority using syntactic analysis to extract semantic relations from Dutch newspaper articles." *Political Analysis* 16(4):428–446.