Plain Text: Transparency in the Acquisition, Analysis, and Access Stages of the Computer-assisted Analysis of Texts

David Romney Brandon M. Stewart Dustin Tingley

May 27, 2015

Text analysis is an exciting tool for social science research, playing an integral part in the rapidly expanding field of computational social science (Grimmer and Stewart, 2013). In political science, research using computer-assisted text analysis techniques has exploded in the last fifteen years, including work studying political ideology (Laver, Benoit, and Garry, 2003), congressional speech (Quinn et al., 2010), representational style (Grimmer, 2010), American foreign policy (Milner and Tingley, 2015), climate change attitudes (Tvinnereim, forthcoming), media (Gentzkow and Shapiro, 2010), Islamic clerics (Nielsen, 2013; Lucas et al., 2015), and treaty making (Spirling, 2011), to name but a few.

While new approaches to text analysis facilitate the use of new kinds of data, they also challenge our traditional approaches to research transparency and replication (King, 2011). These challenges range from new forms of data pre-processing and cleaning, to terms of service for websites which explicitly prohibit the redistribution of their content. Existing references, such as APSA's Statement on Data Access and Research Transparency,[1] provide general guidance. This paper is our attempt to apply these general guidelines to the specific context of text analysis.

We explore the implications of computer-assisted text analysis for data transparency by tracking the three main stages of a research project involving text data: (1) acquisition, where the researcher decides what her corpus of texts will consist of; (2) analysis, to obtain inferences about the research

[1] http://www.dartstatement.org/

question of interest using the texts; and (3) access, where the researcher provides the data or other methods of verification of her results. To be transparent, we must document and account for decisions made at each stage in the research project. Transparency not only plays an essential role in replication (King, 2011) but it also helps to communicate the essential procedures of new methods to the broader research community.

Many transparency issues are not unique to text analysis. There are aspects of acquisition (e.g. random selection), analysis (e.g. outlining model assumptions), and access (e.g. providing replication code) that are important regardless of what is being studied and the method used to study it. We focus here on those issues that are uniquely important for transparency in the context of the variety of text analysis methods available (See Grimmer and Stewart, 2013, for a review of different text analysis methods).

1 Where it all begins: Acquisition

Our first step is to obtain the texts to be analyzed and even this simple task already poses myriad potential transparency issues. Traditionally quantitative political science has been dominated by a relatively small number of stable and publicly available datasets, such as the American National Election Survey (ANES) or the Correlates of War Project (COW). However, a key attraction of new text methods is that they open up the possibility of exploring diverse new types of data, not all of which are as stable and publicly available as ANES or COW. Websites are taken down and their content can change daily; social media websites suspend users regularly. Researchers should strive to record the weblinks of what they scraped (and when they scraped it), so that the data can be verified via the Wayback Machine[2] if necessary and available. This reflects a common theme throughout this piece: full transparency is difficult to impossible in many situations, but researchers should strive to achieve as much transparency as possible.

Sometimes researchers are lucky enough to obtain their data from someone who has already

[2] `http://archive.org/web/`

done the hard work of getting it into a tractable format, usually through the platform of a text analytics company. However, this comes with its own problems—particularly the fact that these data comes with severe restrictions on access to the actual texts themselves and the models used to analyze them, making validation difficult. For this reason, some recommend against using black-box commercial tools that do not provide access to the texts and/or the methods used to analyze them (Grimmer and Stewart, 2013, p. 5). Nevertheless, commercial platforms can provide a valuable source of data that would be too costly for an individual researcher to collect. For instance, Jamal et al. (2015) use Crimson Hexagon, a social media analytics company, to look at historical data from Arabic Twitter that would be difficult and expensive to obtain in other ways.[3] In situations where researchers do obtain their data from such a source, they should clearly outline the restrictions placed on them by their partner as well as document how the text could be obtained by another person working with a company. For example, in the supplementary materials to Jamal et al. (2015) the authors provide extensive detail on the keywords and date ranges used to create their sample.

A final set of concerns at the "acquisition" stage is rooted in the fact that even if texts are taken from a "universe" of cases, that universe is often delimited by a set of keywords that the researcher determined, for example when using Twitter's Streaming API. Determining an appropriate set of keywords is a significant task. First, there are issues with making sure you are capturing all instances of a given word. In English this is hard enough, but in other languages it can be amazingly complicated—something researchers not fluent in the languages they are collecting should keep in mind. For instance, in languages like Arabic and Hebrew you have to take into account gender; plurality; attached articles, prepositions, conjunctions, and other attached words; and alternative spellings for the same word. Over 250 iterations of the Arabic word for "America" were used by Jamal et al. (2015). And that ignores a second concern: the fact that synonyms (e.g. "The United States"), metonyms (e.g. "Washington"), and other associated words ought to be selected on as well. Computer-assisted selection of keywords offer one potential solution to this problem. For

[3] In this case, access to the Crimson Hexagon platform was made possible through the Social Impact Program, which works with academics and non-profits to provide access to the platform. Other work in political science has used this data source, notably King, Pan, and Roberts (2013).

instance, King, Lam, and Roberts (n.d.) have developed an algorithm to help select keywords and demonstrated that it works on English and Chinese data. Such an approach would help supplement researchers' efforts to provide a "universe" of texts related to a particular topic, to protect them from making *ad hoc* keyword selections.

2 The Rubber Hits the Road: Training and Analysis

The researcher should also be transparent about the decisions she makes once her data has been collected. We discuss three areas where important decisions are made: text processing, selection of analysis method, and addressing uncertainty.

Processing

All research involves data cleaning, but research with unstructured data involves more than most. The typical text analysis workflow involves several steps at which data are filtered and cleaned to prepare it for analysis. This involves a a number seemingly innocuous decisions that can potentially be important.[4] Did you stem (i.e. mapping words referring to the same concept to a single root)? Which stemmer did you use? If you scraped your data from a website, did you remove all html tags? Did you remove punctuation and common words (i.e. "stop" words), and did you prune off words that only appear in a few texts? Did you include *bigrams* (word pairs) in your analysis? Although this is a long list of items to consider, each is important and should be mentioned and justified, if not in the paper itself then in an online appendix.

Analysis

Once the texts are acquired and processed, the researcher must choose a method of analysis that matches her research objective. A common goal is to assign documents to a set of categories which are either fixed by the researcher or estimated form the data. There are broadly speaking

[4] See Lucas et al. (2015) for additional discussion.

three approaches available: keyword methods, where categories are based on counts of words in a pre-defined dictionary; supervised methods, where humans classify a set of documents by hand (called the training set) to teach the algorithm how to classify the rest; and unsupervised methods, where the model simultaneously estimates a set of categories and assign texts to them (Grimmer and Stewart, 2013). An important part of transparency is justifying the use of a particular approach and clarifying how the method provides leverage to answer the research question. Each method then in turn entails particular transparency considerations.

In supervised learning and dictionary approaches, the best way to promote transparency is to maintain a clear codebook which documents the procedures for how humans classified the set of documents or words (Hopkins and King, 2010). Some useful questions to consider are:

- How did the researcher determine the words used to define categories in a dictionary approach?
- How are the categories defined and what efforts were taken to ensure inter-coder reliability?
- Was your training set selected randomly?

We offer two specific recommendations for supervised and dictionary methods. First, make publicly available a codebook which documents category definitions and the methods used to ensure inter-coder reliability.[5] Second, documents to be hand coded should be randomly selected unless there is a compelling reason to select training posts based on certain criteria.[6]

When using unsupervised methods, researchers need to provide justifications for a different set of decisions. For example, such models typically require that the analyst specifies the number of topics to be estimated. It is important to note that there is not necessarily a "right" answer here. Having more topics enables a more granular view of the data, but this might not always be appropriate for what the analyst is interested in. Furthermore, as discussed in Roberts, Stewart, and

[5]See Section 5 of Hopkins et al. (2010) for some compact guidance on developing a codebook. A strong applied examples of codebooks are found in the appendix of Stewart and Zhukov (2009) and Jamal et al. (2015).

[6]Random selection of training cases is necessary to ensure that the training set is representative of the joint distribution of features and outcomes (Hand et al., 2006; Hopkins and King, 2010). Occasionally alternate strategies can be necessary in order to maintain efficiency when categories are rare (Taddy, 2013; Koehler-Derrick, Nielsen, and Romney, n.d.); however, these strategies should be explicitly recognized and defended.

Tingley (Forthcoming) topic models may have multiple solutions even for a fixed number of topics.[7] These authors discuss a range of stability checks and numerical initialization options that can be employed.

Uncertainty

The final area of concern is transparency in the appropriate incorporation of uncertainty into our analysis. Text analysis is often used as a measurement instrument, with the estimated categorizations used in a separate regression model to, for example, estimate a causal effect. While these two stage models (measure, then regress) are attractive in simplicity, they often come with no straightforward way to incorporate measurement uncertainty. The Structural Topic Model (STM; Roberts et al., 2014), building on previous work by Treier and Jackman (2008), provides one approach to explicitly building in estimation uncertainty. But regardless of how estimation uncertainty is modeled, in the spirit of transparency it is important to acknowledge the concern.

The second form of uncertainty derives from the research process itself. All of the decisions discussed in this section—decisions about processing the text, determining dictionary words, or determining categories/the number of categories—are often not the product of a single *ex ante* decision. In reality, the process is iterative. New processing procedures are incorporated, new dictionary words or categories are discovered, and more (or fewer) unsupervised topics are chosen based on the results of previous iterations of the model. This is a necessary step in the development of any text analysis model, and we are not advocating that researchers devote themselves an unyielding initial course of action. However, we argue that researchers should clearly state when the process was iterative and which aspects of it were iterative. This documentation of how choices were made, in combination with a codebook clearly documenting those choices, helps to transparently convey any remaining uncertainty in the conceptual framework of the researchers.

[7] Topic models like latent Dirichlet allocation (Blei, Ng, and Jordan, 2003) and the structural topic model (Roberts, Stewart, and Tingley, n.d.) are non-convex optimization problems and thus can have local modes.

3 On the Other Side: Presentation and Data Access

It is at the access stage of a project that we run into text analysis's most common violation of transparency: not providing replication data. A common reason for not providing these data is that doing so would violate the intellectual property rights of the content provider (website, news agency etc.). However, sometimes other reasons prevent the researcher from providing the data, such as the sensitivity of the texts or ethical concerns for the safety of those who produced them. Researchers have found ways around this. Some provide the document term matrix as opposed to the corpus of texts, allowing for a partial replication of the analysis. Others provide code allowing others to obtain the same dataset (either through licensing or web scraping).[8] These are perhaps the best available strategies when providing the raw texts is impossible, but they are each at least partially unsatisfying because the raise the cost of engaging with and replicating the work, which in turn decreases transparency. While intellectual property issues can complicate the creation of replication archives, we should still strive to always provide enough information to replicate a study.

However, there are other post-analysis transparency concerns unique to text analysis that are rarely discussed. We cover two of them here. One of them concerns the presentation of the analysis and results. Greater steps could be taken to allow other researchers, including those less familiar with text analysis or without the computing power to fully replicate the analysis, the opportunity to "replicate" the interpretive aspects of the analysis. As an example, the researcher could set up a browser to let people explore the model and read the documents within the corpus, recreating the classification exercise for those who want to assess the results. With a bit of work, this kind of "transparency through visualization" could form a useful transparency tool.[9]

The second presentation issue relates to level of inference at which research conclusions are drawn. Text analysis is, naturally, done at the text level. However, that is not necessarily the level of interest for the project. Those attempting to use Twitter to measure public opinion, for instance, are

[8] For example, O'Connor, Stewart, and Smith (2013) uses a corpus available from the Linguistic Data Consortium (LDC) which licenses large collections of texts. Although the texts cannot themselves be made publicly available, O'Connor, Stewart, and Smith (2013) provide scripts which perform all the necessary operations on the data as provided by the LDC, meaning that any researcher with access to an institutional membership can replicate the analysis.

[9] One example is the stmBrowser package (Freeman et al., 2015).

not interested in the opinions of the Tweets themselves. But when we present category proportions of texts, that is what we are measuring—such that those who write the most have the loudest "voice" in our results. To address this issue, we recommend that researchers be more transparent about this and either come up with a strategy for aggregating up to the level of interest or justify using texts as the level of analysis.

4 Conclusion

Text analysis is quickly becoming an important part of the social scientists toolkit. Thus, we need to think carefully about the implications of these methods for research transparency. In many ways the concerns we raise here comport with the broader principles of transparent and replicable research (King, 2011). However, as we have highlighted, text analysis produces a number of idiosyncratic challenges for replication from legal restrictions on the dissemination of data to the numerous, seemingly minor, text processing decisions that must be made along the way.

It is clear that there is no substitute for providing all the necessary data and code to fully replicate a study. When full replication is not possible, we should strive to provide as much information as is feasible given the constraints of the data provider. Regardless, we strongly encourage the use of detailed and extensive supplemental appendices which document procedures. As text methods become more entrenched in the field and standard emerge, this may ultimately become less necessary.

Finally, we do want to emphasize the silver lining for transparency. Text analysis of any type requires an interpretive exercise where meaning is ascribed by the analyst to our measure of the text. Through validations of document categories, and new developments in visualization, we are hopeful that this interpretive exercise can be exposed to the reader as well. Providing our readers the ability to not only evaluate our data and models, but also to make their own judgments and interpretations, is the fullest realization of research transparency.

References

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent dirichlet allocation." *the Journal of machine Learning research* 3: 993–1022.

Freeman, Michael, Jason Chuang, Molly Roberts, Brandon Stewart, and Dustin Tingley. 2015. *stmBrowser: An R Package for the Structural Topic Model Browser*.

Gentzkow, Matthew, and Jesse M Shapiro. 2010. "What drives media slant? Evidence from US daily newspapers." *Econometrica* 78 (1): 35–71.

Grimmer, Justin. 2010. "A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases." *Political Analysis* 18 (December): 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 (January): 1–31.

Hand, David J et al. 2006. "Classifier technology and the illusion of progress." *Statistical science* 21 (1): 1–14.

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

Hopkins, Daniel, Gary King, Matthew Knowles, and Steven Melendez. 2010. "ReadMe: Software for automated content analysis." *Institute for Quantitative Social Science*.

Jamal, Amaney A., Robert O. Keohane, David Romney, and Dustin Tingley. 2015. "Anti-Americanism and anti-interventionism in Arabic Twitter discourses." *Perspectives on Politics* 13 (March): 55–73.

King, Gary. 2011. "Ensuring the Data Rich Future of the Social Sciences." *Science* 331 (2011): 719-721.

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): 326–343.

King, Gary, Patrick Lam, and Margaret E. Roberts. N.d. "Computer-assisted keyword and document set discovery from unstructured text." Working paper.

Koehler-Derrick, Gabriel, Rich Nielsen, and David A. Romney. N.d. "The lies of others: Conspiracy theories and selective censorship in state media in the Middle East." Working paper.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting policy positions from political texts using words as data." *American Political Science Review* 97 (02): 311–331.

Lucas, Christopher, Richard Nielsen, Margaret Roberts, Brandon Stewart, Alex Storer, and Dustin Tingley. 2015. "Computer assisted text analysis for comparative politics." *Political Analysis* 23: 254–277.

Milner, Helen V, and Dustin H Tingley. 2015. *Sailing the water's edge: domestic politics and American foreign policy*. Princeton University Press.

Nielsen, Richard. 2013. "The Lonely Jihadist: Weak Networks and the Radicalization of Muslim Clerics." Ph.D. diss. Harvard University. Ann Arbor: ProQuest/UMI. (Publication No. 3567018).

O'Connor, Brendan, Brandon M Stewart, and Noah A Smith. 2013. "Learning to Extract International Relations from Political Context." In *ACL*. pp. 1094–1104.

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): 209–228.

Roberts, Margaret, Brandon Stewart, and Dustin Tingley. Forthcoming. *Navigating the Local Modes of Big Data: The Case of Topic Models*. New York: Cambridge University Press.

Roberts, Margaret, Brandon Stewart, and Dustin Tingley. N.d. "stm: R Package for Structural Topic Models." Working Paper (https://github.com/bstewart/stm/blob/master/vignettes/stmVignette.pdf?raw=true).

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 (October): 1064–1082.

Spirling, Arthur. 2011. "US Treaty Making with American Indians: Institutional Change and Relative Power, 1784–1911." *American Journal of Political Science* 56: 84-97.

Stewart, Brandon M, and Yuri M Zhukov. 2009. "Use of force and civil–military relations in Russia: an automated content analysis." *Small Wars & Insurgencies* 20 (2): 319–343.

Taddy, Matt. 2013. "Measuring Political Sentiment on Twitter: Factor Optimal Design for Multinomial Inverse Regression." *Technometrics* 55 (4): 415–425.

Treier, Shawn, and Simon Jackman. 2008. "Democracy as a latent variable." *American Journal of Political Science* 52 (1): 201–217.

Tvinnereim, Endre. forthcoming. "Explaining topic prevalence in answers to open-ended survey questions about climate change." *Nature Climate Change*.