

Here's how text analysis is transforming social-science research

By **Joshua Tucker** May 27, 2016

The journal *Political Analysis* has recently published a “[virtual issue](#)” on “Recent Innovations in Text Analysis for Social Science.” In addition to the guest editor’s introduction, there are seven papers in the virtual issue. All of the papers are available for free reading [online](#), for a limited time. I spoke to University of California at San Diego political scientist [Margaret Roberts](#), who edited the issue, about the subject matter. What follows is a lightly edited version of our discussion.

Joshua Tucker (JT): What exactly is the field of “text analysis,” and how is it different from simply reading existing texts?

Margaret Roberts (MR): “Text analysis” is revolutionizing social-science research. [Automated text analysis](#) allows social scientists to use computational methods to quickly study the content of huge numbers of political documents. Humans write billions of words every day about their social lives and about politics. People share their political opinions in social-media posts, governments record the minutes of meetings and the text of legislation, and newspapers recount political events in daily publications.

We are so prolific that social scientists could never read every document that contained information about their topic of study — doing so would take lifetimes of doing nothing other than reading these texts!

However, new methods allow us to represent this text quantitatively. This allows analysts to use statistics to summarize the content of the documents, condensing years of reading into minutes of computation. Text analysis quickly sums up entire corpora of text, allowing social scientists to systematically study politics encoded in the text, from [measuring political influence on Twitter](#) or [uncovering what the Chinese government is censoring online](#).

Text analysis can also be used to assist reading: It can flag a selection of documents that should be read in more detail, focusing social scientists on important, representative or influential texts.

JT: How are political scientists using text analysis?

MR: Automated text analysis has allowed researchers to analyze political phenomena at a previously impossible scale.

Just in this virtual issue, researchers have used these methods to analyze thousands of Senate press releases to describe how politicians in Washington explain their political decisions to their constituents, used millions of newspaper articles to pinpoint instances of militarized disputes between countries and analyzed thousands of Islamic legal rulings to better understand which religious topics are more frequently discussed by jihadist clerics.

Social scientists are using text analysis methods in existing research designs. For example, we've created methods to summarize open-ended questions in surveys and have used text analysis methods to analyze text produced by online experiments.

Some of these political data are brand new and some of them have existed for decades, but we are only now unlocking their potential to obtain better descriptions of important phenomena from politics to religion to political opinion.

JT: What are the major challenges facing more widespread adoption of text analysis in research?

MR: Currently, the challenge for text analysis isn't getting data. Large amounts of text data are being produced and documented online at a rate faster than social scientists can use them.

The most important challenge is being able to estimate concepts that are of interest to social scientists directly from the texts. "Big" data and in particular text data are only as useful as the methods we have to use them to answer questions. Social scientists would like not only to automatically extract measures of topics and sentiment from texts, but also to uncover more complex social phenomena such as persuasion, humor, sarcasm, innovation and influence.

Researchers are making progress developing statistical methods that can summarize the complex social processes that are richly reflected in text, and the authors in our special issue make significant strides toward this goal with the methods that they develop for automated text analysis. Statisticians, social scientists, companies and computer scientists are making these methods available through statistical software so others can use them off the shelf (see some recent software by political scientists).


JT: What was the most interesting thing you learned when putting together this special issue? What are the most exciting questions that the articles in the special issue could be used to answer in the future?

MR: It used to be that we could only study people by meeting them in person, such as through surveys or interviews conducted by painstaking travel, sometimes to remote locations. This severely constrained the types of people we could reach and study.

But now, people around the world are posting political information online from places social scientists had trouble reaching before — from the middle of large-scale protests to the center of conflict zones. And since automated text analysis can be applied to any language, we can use it to study politics around the globe. This issue contains legislative debates in Ireland, voter registers in Kenya, Chinese- and Arabic-language social-media posts and political manifestos from countries all over Europe. The articles in this issue deal with texts in multiple languages and even across languages.

JT: How long will these articles be available for public access?

MR: The virtual issue will be online and freely available until September 2016.

Joshua Tucker is a Professor of Politics at New York University. He specializes in voting, partisanship, public opinion, and protest, as well as the relationship of social media usage to all of these forms of behavior, with a focus on Eastern Europe and the former Soviet Union.  Follow @j_a_tucker

