Chapter 2: Social Science Research and Text Analysis

Justin Grimmer* Margaret E. Roberts † Brandon M. Stewart ‡

January 8, 2018

Abstract

Incomplete and preliminary. Do not distribute or cite.

Texts are increasingly used to make social science inference, because they are often the medium where social science occurs. The abundance of texts have, however, presented a challenge to social scientists. At the same time an explosion of new technology has provided social scientists with the ability to utilize much larger data sets. Yet, many of these techniques are developed in other fields or are often developed with other purposes in mind.

In this chapter we provide a framework and a general set of principles for utilizing text as data methods in social science research. Our framework applies generally across ways scholars might utilize text in their research. As we will explain throughout the book, it covers methods that are traditionally known as "machine learning" methods, models that come from statistical approaches, and even qualitative approaches to texts that involve mere hand coding. The reason our framework is general is because it focuses on how social scientists use texts in their research, rather than focusing on the tools that are deployed to accomplish those goals.

With a collection of texts, we explain how text as data methods can contribute to inferences at three stages of the research process that we identified in the last chapter: discovery, measurement and causal inference or prediction. At the *discovery* – or hypothesis generation – phase, text methods can help illuminate new concepts, new methods of organizing data, or suggest insights that deserve further investigation. Text as data methods provide researchers with new ways to organize their texts—including identifying clusters in the data, an underlying spectrum, or words that charcterize a particular group of people. This new organization can prompt social scientists to read the texts differently and draw connections that they otherwise would have missed.

With concepts in hand–either from a discovery stage, from a theoretical model, or other intuition—scholars often want to *measure* or describe the prevalence of particular concepts in their data or to characterize where individuals and texts align on a spectrum. For example,

*Associate Professor, Department of Political Science, University of Chicago
†Assistant Professor, Department of Political Science, University of California at San Diego
‡Assistant Professor, Department of Sociology, Princeton University

Table 1: Key Principles About Text as Data and Social Science Research

1) Social Science Inferences are Necessarily Sequential
2) The codebook function, g, is central: Text as data methods are about compression
3) There is no general theory of language, nor globally best method.
4) Text as Data Methods Do Not Replace Humans, They Augment Them
5) Validate,Validate,Validate

we might be interested in learning the amount of legislation that falls within a broad range of policy agendas (POLICY AGENDAS PROJECT) and the comparative manifesto project measures the prevalence of topics in party manifestos across the world. Numerous other papers, books, and dissertations read documents, and assign them to categories by hand. The increasing prevalence of text data and the preponderance of methods to do the measurement has lead to an explosion in interest in measuring quantities in texts and the ability to measure quantities in larger collections of texts. The prevalence of text means that the measures are often granular, providing insights into behavior otherwise difficult to detect. Measurement is useful on its own. Description can provide valuable summaries of the data that may inform theories, provide the measures necessary for causal inferences, and characterize the state of the world. To do this, researchers have to demonstrate that their method of measurement does indeed describe the concept or behavior they would like to measure –they have to validate it.

Once a concept is discovered and we have measures, researchers are able to use those measures from text data to make a *causal inference*, or learn the effect of some intervention, or make a *prediction*, what will happen in the future, given the past. For example, researchers might assess the causal effect of a president's speech on the salience of news coverage about a particular topic or they may be interested in how political content affects users' engagement in online forums. Or, researchers may wonder the extent to which information in text can predict events in the world, like stock market movement or conflict (CITATIONS). The methods for discovery, measurement, and testing using text data vary not only because they are focused on different stages of the research process, but also vary in how stringently they rely upon the specific assumptions of the model.

Discovery, measurement, and causal inference are separate inferential goals. In spite of the differences in inferential goals at each stage of the research process, we advocate for five key principles when using the methods for social science research. Table **??** introduces our five principles, which we develop in depth below. Our first principle is that social science inferences are necessarily sequential. We learn about the world through an iterative process that involves exploration, measurement, and testing. Our second principle is that the codebook function g is central, because text as data methods are about compression. When applying text as data methods, the general goal is to reduce the complexity of the text to find some interesting features, rules for classification, or some new view of the data. Given this goal, our efforts center around developing procedures for compressing information and evaluating the content of the compression. Our third principle is that there is no general theory of language and, as a result, no globally best method for analyzing text methods.

Not only do we have different social scientific goals, text-based methods often have different statistical goals. The result is that there is no one method that will accomplish all tasks and there is no one method that strictly dominates all other methods as a text analytic tool. Rather, different tasks require different techniques. Our fourth principle is that text as data methods do not replace humans, they augment them. We will see that the methods we introduce in this book make humans more effective readers, lower the cost of analysis, and in many ways change the way we read. But applying text methods is still a fundamentally qualitative research activity—and as a result, humans will need to read, interpret, and explain the output of the methods. And our fifth and final principle is Validate, Validate, Validate. The methods we describe in this book all make drastic assumptions about how texts are created and then deployed. For example, in Chapter 3 we describe how we discard word order to create text for analyzing quantitatively. In Chapter 4 we describe methods that make unrealistic assumptions about how texts are produced. And across the many methods that we present in this book (and methods that are left out) there are no theorems that show that the methods are guaranteed to capture the most important features of text. Rather, experience has shown that the methods tend to be useful in many settings. But to assess the utility in any one setting requires extensive validation.

In this chapter we describe our approach to social science inference and our principles for applying text methods. To provide more context for each stage in the research process we first offer a description of one research program: an analysis of how **?** discovered, measured, and then confirmed the process of Chinese censorship.

1 Discovering, Measuring, and Causal Inference: How the Chinese Government Censors Posts

Ideas and research programs rarely emerge as a straightforward product, even though they are often reported that way in papers. While research papers often portray thoughtful questions and careful hypotheses as emerging after careful non-data driven contemplation and observation about the world, the reality is that those questions and hypotheses often emerge from an initial inspection of data. This inspection is usually guided by a substantial background knowledge and preexisting theories of how the world works which informs the discovery of a theoretical tension or empirical puzzle. This is perhaps most clear when working with textual data. Often times, research projects begin with a particular goal and then shift to explain an interesting aspect of the data. Other times, research projects discover some previously unknown feature of the data—and then the researchers change focus and seek to explain that new and interesting feature.

The non-traditional approach is evident in **?**, who develop a measure of censorship in China (KPR hereafter). KPR initially sought to use blog posts as a measure of public opinion in China— an important objective because of the difficulty of running surveys in an authoritarian environment. To this end, KPR downloaded millions of posts from over 1,400 social media websites. After making notes on a few potentially controversial posts

and logging the urls of the posts, the researchers went back to the original posts to better understand the post's context. When they returned to the posts, however, they noticed something surprising: many of the posts were now missing. Rather than the original social media posts, KPR now found pages that proclaimed that the content had been removed – an indication of government censorship.

KPR had accidentally stumbled upon a research design that would enable them to directly measure the rate of Chinese censorship—they had used the exploration of their data to both *discover their question of interest*. They had also discovered one conceptualization or way to organize the texts: they could view the social media posts as censored or not. The identification of this particular empirical phenomenon – the removal of documents from the web – derives its significance by the researchers embedding it into the broader social science context of censorship in authoritarian regimes (CITATIONS HERE). It should also be clear that after the fact it is obvious that some social media posts are censored in China. But before hand it was far from obvious that collecting social media posts and then revisiting them could provide a valid design for studying Chinese censorship.

This initial exploration of their data led them to generate a hypothesis about why certain texts would be censored. At this point they are ultimately interested in answering a causal inference question, but do not yet have a viable potential intervention that could explain the censorship decision. To identify this potential intervention, a computer-assisted manual examination of the text gave them the impression that censorship rates tended to be much higher when the posts were about potential or existing collective action events in China—when groups of people would potentially come together, which could lead to a protest of the Chinese government. This suggested a new conceptualization that suggested a key quantity to measure: to what extent does a post make a reference to a collective action event?

Critically, this exploration stage gave them reason to believe that other conceptualizations would be less useful for answering the causal question. For example, sentiment, whether the post was critical or supportive of the government, did not seem to explain censorship decisions. Thus, using their texts, KPR were able to *discover* two more conceptualizations of the texts: the events the posts related to and whether the posts were critical of the government, neutral, or supportive of the government. If they were right, then the decision to censor a text would be unrelated to whether the post was critical about the government or not, but would focus on posts that described protest events.

Rather than a process detached from their texts, KPR used both a subset of their data and statistical methods to discover their question of interest, generate a hypothesis, and to formulate the implied conceptualization. We will describe a general process for this method of *discovery* in Chapter 4. Given this information, they then set out to *measure* the texts according to their conceptualizations. To measure the topics of the posts, they used a supervised learning method, based on the keywords of the posts. KPR identified an initial set of keywords that they hypothesized would allow them to identify whether the posts were about a particular collective action event. They then iteratively refined that list to ensure the keywords captured only collective action events and were neither too broad nor too narrow. Measuring whether a post was censored was straightforward, based on their own

records of the post. It was less straightforward, however, to measure the post's sentiment toward the government. As often happens with social science questions, their goal was not to characterize each individual document's sentiment, it was to characterize the distribution of sentiment across censored and uncensored posts. For this task, they used a distinct supervised learning algorithm—ReadMe, which we cover in Chapter 5—to measure whether censored posts were supportive or critical of the government. Specifically, KPR sampled posts and hand labeled them as critical or supportive of the regime. They then used ReadMe (**?**) to extrapolate from the hand-labeled documents to the entire collection of posts and measure the proportion of documents that were supportive or not of the government.

Using an initial subset of posts, then, KPR were able to discover a theoretically interesting research question. They then moved to a larger data set, where KPR were able to measure the key quantities of interest. With the measurement in hand, KPR are able to estimate a social science causal effect of interest: the average effect of a post being about collective action on the probability that a post is censored. They also attempt to estimate the average effect of senatiment on the probability of censorship. While the measures necessary to answer those questions are now in hand, the measures alone are not sufficient for answering the causal inference questions. A design is also necessary to eliminate confounding: other factors that are correlated with either collective action potential or sentiment and the outcome, but are distinct. For example, if collective action social media messages tend to be issued from people who are censored at a high rate we will confuse the message for the person.

To test the hypothesis, KPR utilize exogenous events that occur in China, which allows them to examine the government response. They find that posts around collective action events are censored at an extremely high rate—strong evidence that the government is censoring posts about events that might cause the public to come together. They also find that the government's decision to censor posts or not is essentially unrelated to whether the post supports the government. Their conclusions provide evidence that the government's censorship rules are not designed to merely suppress dissent against the regime. It also provides an example of how text can be used to make a causal inference. In Chapter 6 we describe a general approach for causal inferences with texts, including when text is the *intervention*.

In their first article, KPR make a compelling case based on observational data that censorship focuses on three things: collective action, pornography and criticism of censors. Their data was close to the real world but did not allow for the possibility of manipulating the subject of the post directly. The ability to randomly assign the variable of interest (in this case the topic of the post) is a key component of the most rigorous causal inference designs. Having established their hypothesis in **?**, KPR produced a second study **?** that used a randomized field experiment to verify that their theory of censorship was true. In this follow up study, they created accounts on one hundred social media sites across China and submitted text to these sites, randomly assigning the text to discuss protest events or events that were not related to protests and randomly varying whether the post was critical or supportive of the government. These experimental results produced the same conclusion as the observational results, providing further verification that protest-related topics were causing censorship.

KPR's ground breaking insights into the Chinese government did not begin with a well-stated question of social science causal inference. But this does not mean that their research design is over-fit or atheoretical. The initial exploration of the data was based on a subset of the data. This ensured that KPR would have the opportunity to demonstrate that they were wrong about how the Chinese government censors data on new data where a different pattern might maintain. Theory also has an important role throughout KPR's design. They began with baseline theoretical accounts of how the Chinese government censors posts and using that theory lead them to pose new questions from their data. Their initial work inspired a second study which sought to reinforce claims made through observational data with data collected in a more controlled laboratory-like environment.

While not explicitly designed to test a formal model, the results of KPR's study has also inspired new formal theoretic models that seek to explain and contextualize KPR's findings. In this sense, we see the specific instantiation of our more general point that science is sequential and collaborative.

Using KPR's experience as a reference, in this chapter we describe the research process when using text as data. While we focus on decisions that must be made when working with text as data, our argument is more general and illuminates one model of how empirical social science can proceed. It is also worth noting that scholars must enter into the research process at various stages. In the conclusion to this chapter, we explain how our framework can be used if scholars already have well defined questions or have key measures of quantities of interest in hand.

2 Social Science Inferences Are Necessarily Sequential

The goal of social science is to build robust theoretical explanations for social phenomena. But theory building is rarely simply about accumulating evidence from theoretical deductions and then revising the theory accordingly. This focus is a product of a time when acquiring data, running surveys, conducting experiments, and interviewing subjects was expensive and thinking was relatively cheap. The cost of thinking has stayed relatively constant overtime, but the cost of data has plummeted. The result is that it is practical for social scientists—even those on the smallest of budgets—are able to include new data acquisition as an explicit part of the research process. This means that initial samples of data are viable to be used as an exploration to refine the research question—and to potentially stumble upon new and useful questions. The new ways to organize the world suggest new measurements, but then also new causal inference questions. This leads us to pursue new measures and research designs, which we then use to estimate causal inferences. Or, we use the new measures to make predictions about what will happen in the future.

Rather than the sequential approach to social science being atheoretical, each stage of the research process is both informed by theory and can help us build new theoretical explanations and then test those explanations. For example, developing a conceptualization requires knowing how prior theoretical work organized the world and the extent to which a new organization is actually new. Measures are interesting in so how much as they corre-

spond to some theoretical quantity of interest (or several quantities of interest under different theories). The estimated causal effects can help us to build new theories and to test existing theories. Causal effects might help us to better understand and determine the extent to which observable implications are found in a data set and which observable implications are not present.[1]

Learning from data and updating theories and then testing those revised theories is not a strong move away from the usual social science process. Rather, we view it as a more honest account of how social science research tends to be done. We think this is useful, because the increased transparency will also make clear the value of different methods at particular phases of the research process. For example, we describe discovery methods in Chapter 4 of this book, which tend to get little attention in the social science. We think this is due, in part, to the tendency to pretend that conceptualizations are already known before the real research begins.

As we will emphasize throughout the book, the sequential nature of research leads us to regularly recommend that analysts split their sample. At nearly every stage of the research process—whether discovery, measurement, causal inference, or prediction—we encourage the division of our data into a training and test set. In discovery, this is essential to learn whether a particular organization of texts is only prominent in one subsample, or if the prevalence is found in other data sets. In measurement the use of a split sample is more traditional and ensures that our functions that are used to classify objects are not overfit and the fresh data ensures that we can accurately evaluate the performance of our classifier on new data. In Chapter 6 we show that the training/test split is essential for using machine learning methods in causal inference. The training stage enables us to tune our method, discovering conceptualizations that are likely to give rise to interventions that exert a causal effect on some outcome of interest. The test stage enables us to credibly evaluate the size of the causal effect, while also avoiding some of the common problems with fishing in causal inference. As we will explain, the train/test split solves problems that other mechanisms have been developed to do, like the pre-analysis plan. The benefit of having a train/test split is that it enables the researcher to have an explicit discovery phase, which is often ruled out in preanalysis plans. Further, pre-analysis plans are often only effective when others are present to regulate their content, which if often lacking.

The recommendation to split samples is usually met with two different kinds of objections. The first objection is that a particular experiment might be very costly and the decrease in power to split the sample is not justified. As we explain in more detail in Chapter 6, this objection is intuitive, but our intuition is exactly wrong. Especially in instances where stakes are high, interventions are expensive, and the consequences for public policy are clear,

[1]Importantly, the theory building can be with formal models, which are an integral feature of social science but one that is largely outside the scope of this book. Formal models can be used within the research process in order to derive hypotheses and then to test. Formal models can also be built on the output of models, helping researchers to interpret the conclusions of research. But crucially formal models can also help us to interpret the findings from a study. By thinking carefully about actors' incentives when responding to intervention, we can better insights into what an effect tells us about the behavior of individuals and how it reveals patterns of strategic interaction. (CITE BDM BOOK HERE)

we want to have the most confidence in the effects we estimate. A split sample provides a robust guard against fishing, while also ensuring our data cleaning rules take into account the realities of a particular data set. The second objection is that not some historical data may occur only once and therefore splitting a sample can happen only once. In many ways this is true, some events only occur once and therefore provide us with only one data set (and therefore one sample split). But as Fowler and Montagnes (2018) suggest, there are often analogous interventions that could be studied. For example, they examine the compelling finding in Malhotra (XXX) that college football scores affect Congressional elections, finding no indication that National Football League (NFL) games affect the outcomes.

3 The Codebook Function, g, is Central: Text as Data Methods are about Compression

Text is inherently high-dimensional. To get a better sense of what this means, consider one of the greatest speeches in American political history. On the eve of his assassination, in the context of a large number of death threats, Martin Luther King, Jr. delivered a speech entitled "I've Been to the Mountaintop" in Memphis, Tennessee. In the rousing speech where King confronts the threats on his life directly, King closes with a prophetic declaration:

> Like anybody, I would like to live a long life. Longevity has its place. But I'm not concerned about that now. I just want to do God's will. And He's allowed me to go up to the mountain. And I've looked over. And I've seen the Promised Land. I may not get there with you. But I want you to know tonight, that we, as a people, will get to the promised land!

In many ways, "I've Been to the Mountaintop" is high-dimensional, like other pieces of text. It is high-dimensional in the obvious way—language depends on the order that words are written. In this sense, the speech is unique, only taking on the particular meaning because of the order of the words. And requires this exact sequence of words to convey the exact same meaning. Beyond the order dependence of words, however, interpretation of the speech depends on context, adding other dimensions to the speech. Part of the speech's power comes from when it was delivered and who delivered it: on the eve of a great American's assassination, an iconic civil rights leader discusses his vision of the future. And the speech is also powerful because it makes a strong reference to biblical stories, in particular the book of Exodus. King is implicitly comparing himself to Moses, the black audience members to the enslaved people escaping Egypt, and the quest for equal rights and justice to the escaped slaves finding a land of their own.

The goal of text as data methods is to develop a *codebook* function that reduces the high-dimensional information in the text to a much lower-dimensional representation that is then used for social science research. The reason for reducing the information are several. First, reducing information can make the text interpretable in a way that would be otherwise difficult. The reduction in dimensionality is a major reason that text as data methods

are useful for discovery. Reducing the complexity of text in discovery often provides an organization of the documents that then informs the way we read those documents. It can also cause us to rethink how we view a particular phenomenon, leading to new questions, ideas, organizations, and research projects. When performing a measurement, the reduction of information, enables us to understand what we are measuring and why it might be relevant as a quantity of interest. For example, reducing political text and roll call voting decisions to a single dimension can be incredibly useful for summarizing political decisions. The hope when doing reduction for interpretation is that we preserve the useful and interesting facets of the texts.

A second reason for reducing the dimensionality of text is for statistical properties. When we are attempting to use texts to make predictions, infer how texts affected an individual, or trying to infer how texts are similar, we are unable to work with the entire text because there is simply too much information. This curse of dimensionality can manifest in several statistical applications of text as data methods. Even with the biggest data sets we are simply unable to learn about how all the complicated features and dimensions of a text apply to a particular problem. If we fail to reduce the dimensionality when attempting to make predictions we will overfit and our forecasts will perform poorly. If we do not reduce the dimensionality when attempting to infer the causal effects of a speech we will lack the basic information necessary to reliably estimate causal effects of interest. The only way to proceed, then, is to reduce the information in the text.

A third reason to focus on reducing the dimensionality of text is that most social science theories are about relatively low-dimensional and/or a small number of categories. For example, political economy theories of politics often suppose that conflict happens along a low-dimensional ideological spectrum. In order to test the observable implications from these theories, then, we need measures of individuals along the dimension. Or, we might be interested in describing the nature of campaign advertisements. This literature tends to create a dichotomy between negative and non-negative advertisements, with more refinement of both categories sometimes used.

Given its importance, we will focus our efforts in this book on understanding the particular g function we are estimating, what properties it might have, and how we know if we have done a good or bad job. The function will take on several forms. For example, we might attempt to obtain a g function that will take a particular document and assign it to a known category. Or, we might want a function that will take a collection of documents, partition them into a set of categories and then assign documents to categories. We might use a g function to discover the treatments that are present in a collection of documents. Or, we might use research assistants to hand classify documents into a set of categories, constituting a manual g function.

As we explain below, the g function's overall goal is to retain the most important content of the documents, while discarding the features of the document that are irrelevant for the analysis at hand. This is why we refer to the g function as a distillation, or summary, of the text. We are not attempting to provide a comprehensive account of what is in the text and all facets of meaning. Rather, we want to retain the most relevant content. This objective is

intentionally vague. Further refinement of the goal requires that we know more about what we are trying to accomplish and where we are in the research process. We discuss this in the next section.

4 There is No General Theory of Language Nor Globally Best Method For Text Analysis

While we have an overarching goal of distilling texts into the most useful content, we will not introduce a general theory of language to achieve this summary. Nor will we introduce a single method for reducing the information in the texts. We do not introduce a general theory or model of language because there is not a model available that would be generally useful for social scientific tasks.

We do not introduce a general theory of language because we are unaware of one that would be useful for the ways social scientists apply text data in their research. Indeed, across the field of computational linguistics there is not a comprehensive model of language that is generally applied to comprehend text. The lack of a model of language is due, in large part, to the complexity of language in the world. As we discussed above, the meaning of language is found not just in the words—but also in the order they are spoken, the characteristic of the speaker, and the social context of the reader. This means that any attempt to comprehensively distill the work down to a general model of language generation, or even a general model for a specific language, will be fraught with difficulties. And will likely be impossible, or so difficult as to not be useful.

We also lack a comprehensive model of language because the features of language that are of interest, the content that we would like to model, depends on what we want to learn from the text. For example, the content that we would want to extract from a text is qualitatively different if we are interested in learning the topic of discussion than the content we would extract if we are interested in learning the sentiment from a text. And we are interested in fundamentally different things when we attempt to learn the ideology of a speaker as opposed to trying to identify who wrote a text (a task from the field of stylometry).

In short, we do not have a comprehensive model of language both because it is difficult to develop and because the relevant features of text to model depends on our task. The lack of a theory also implies that there is no generally best method for analyzing text. The lack of a general method is because of the multitude of ways that we might use text data. Discovering an organization of texts is fundamentally different than classifying documents into preexisting categories, which is different than inferring the causal effect of a text.

The methods that are applied to text as data also tend to lack the statistical theory found in many other areas of statistics. Yes, there are important and deep statistical theoretical properties for the estimators that we will discuss in this book. And those properties are important for us to know how to evaluate the content. But, not even the most theoretically well developed methods have theorems that relate the performance of the method back to natural language as it is spoken. Rather, the properties tend to be developed conditioning

on already

HOW DO THE METHODS WORK, THEN? It is not based on theorems.

But seemingly disparate methods have surprising connections. We emphasize those connections with a universal notation throughout our book. In particular, we adopt the following conventions:

Notational conventions.

We will be self-conscious about the connections and will try to draw out comparisons

All the methods will be wrong, some will be useful.

5 Text as Data Methods Do Not Replace Humans, They Augment Them

Key thing to argue here is that text as data methods bridge the quantitative qualitative divide. Indeed, we're engaged in a fundamentally qualitative task: learning from texts. This will require intense reading of the texts. As we will see throughout the book, text as data methods provide an augmentation of the kind of approaches we usually take with text.

One way they augment is that they can make some activities cheaper. This is supervised learning. And scaling.

The methods also help us to identify organizations that we otherwise might miss. This includes conceptualizations and low-dimensional representations. The key here is that text as data is not equivalent to doing big data analysis. To see why consider conceptualizations with merely 100 documents. There are a lot of ways to conceptualize that. Many more than a computer could consider, but substantially more than a person ever would. The notion that we would be able to parse those is a fool's errand.

What is obvious, then, is that quantitative methods and qualitative methods have a lot to contribute.

This may mean that we are able to identify causal effects that we otherwise might. Explain how the discovery helps with causal inference.

6 Validate, Validate, Validate

The ways we evaluate the g function will depend upon the particular task we are using the distillation to accomplish. If we are attempting to *discover* some underlying organization or conceptualization then the primary evaluation will be what the g function suggests and whether this provides us with useful insights. Though, we will also want to evaluate the g function to ensure that we are correctly interpreting the function. When used to *measure* according to some conceptualization, we can ask about the g function's bias and its precision. When used to make a *causal inference* we will also look for evidence that the g function conforms with assumptions that are necessary to avoid bias in our estimates. And when used to make predictions we will develop schemes for asking