Memo: Machine readable text and the scientific study of diplomacy

Leah Windsor *  Mark Nieman †  Zuhaib Mahmood ‡
University of Memphis  Iowa State University  Michigan State University

October 2018

Abstract

The use of textual data has seen a burgeoning interest from scholars over the past several decades. From inferring political party positions to measuring changes in the dimensionality of human rights, examining the public and private motivations of actors to looking at diplomatic communications, exploring the growing computational and statistical resources have led to the opening of new frontiers in political research (Slapin and Proksch, 2008a; Greene, Park and Colaresi, forthcoming). Concurrently with the availability of these data sources, Political Scientists have benefited from increasingly sophisticated and careful techniques to both process and extract information from these texts (Monroe, Colaresi and Quinn, 2008; Grimmer and Stewart, 2013). The goal of this memo is to complement the increasing availability of methodological roadmaps with a similar roadmap for the prospects, pitfalls, and limitations of machine-readable text for research. A secondary goal of this memo is to provide a starting point for bridging between theories of diplomacy and potential empirical datasources to test these theories.

*Leah.Windsor@memphis.edu
†mdnieman@iastate.edu
‡mahmoo21@msu.edu
The use of textual data has seen a burgeoning interest from scholars over the past several decades. From inferring political party positions to measuring changes in the dimensionality of human rights, examining the public and private motivations of actors to looking at diplomatic communications, exploring the growing computational and statistical resources have led to the opening of new frontiers in political research (Slapin and Proksch, 2008a; Greene, Park and Colaresi, forthcoming). Concurrently with the availability of these data sources, Political Scientists have benefited from increasingly sophisticated and careful techniques to both process and extract information from these texts (Monroe, Colaresi and Quinn, 2008; Grimmer and Stewart, 2013).

Given the potential benefits of using text-as-data in the study of diplomacy, why is it not more frequently used? There are a number of difficulties with examining textual data. For one, it can difficult to identify the appropriate documents or data appropriate for the research question. It may also be difficult to learn and apply the appropriate methodological tool. Perhaps the greatest roadblock is that many diplomatic communications are secret (or remain hidden), are not in usable formats, or are behind paywalls and other impediments (Gill and Spirling, 2015). The goal of this memo is to complement the increasing availability of methodological roadmaps with a similar roadmap for the prospects, pitfalls, and limitations of machine-readable text for research. A secondary goal of this memo is to provide a starting point for bridging between theories of diplomacy and potential empirical datasources to test these theories.

Types of text data and their relevance to the study of diplomacy

The natural starting point for reviewing textual data sources is to understand the substantive types of textual data which are applicable to the field. For example, the most obvious applicable source of data is that of speech: the verbal communications of political actors to one another, to a particular audience, or to the general public. This type of data has direct application to the mechanisms underlying political communication. For instance, an application to domestic politics by Proksch and Slapin (2012) tests whether political incentives derived from institutional structure—in their case, whether the electoral incentives of individuals vs. the broader party leads to differential outcomes in public communication. The underlying question of interest is to test how political actors respond to the tension between party interests and individual interests; the use of speech here is a measurement decision, whereby this type of text provides a fine-grained measurement at the individual level which has the capacity to actually capture this variation. For example, an immediate application to international politics would be as a measurement for whether material power makes deviation from a major-power position in hierarchical relationships (Lake, 2009) more pronounced or likely.

Speech as data also has applications for bringing empirical leverage over theories which were heretofore relatively abstracted, often in the form of game theory. For instance, McManus (2014) uses automated content analysis—specifically a dictionary-based measure—to classify statements made in the context of a Militarized Interstate Dispute by U.S. presidents by their level of resolve. She shows empirically that as is theoretically expected, higher levels of stated resolve are strongly correlated with favorable outcomes for the United States; she also shows that the highest levels of
resolve are relatively rare, in these statements, opening up avenues for further work. Of course, as McManus also notes, this test is still subject to questions of which direction the causal arrow runs: here, we see an example underscoring the fact that while these textual sources of data are useful and interesting, they do not supersede careful theorizing in answering important questions. Another use for political speech data is estimating leadership tenure and survival through political crises. Language can provide insight into the regime’s perception of the socio-political instability, as well as the crisis management strategy. For example, positively-valanced language improves public opinion ratings (Love and Windsor, 2018), and it also is correlated with leaders remaining in office through social upheaval and political turmoil (Windsor et al., 2018).

Another line of inquiry into the language of political leaders, called Leadership Trait Analysis, uses computational content analysis on extemporaneous political speech (Hermann, 2005). This methodology has been used to explain phenomena such as the foreign policy behavior of leaders (Hermann, 1980), conceptual framework of central bankers and the banking industry (Thies, 2006), and gendered dynamics of diplomacy (Kesgin, 2012). Role theory (Sarbin and Allen, 1954) and operational code analysis are other well-established approaches to understanding the foreign policy preferences of international actors, such as Putin (Dyson and Parent, 2018).

There are interesting areas for leveraging the data beyond the text itself. Analysis of multimodal communication, including nonverbal, audiovisual, and vocal cues, can reveal a more complex picture of how different channels convey congruent or dissonant signals (Windsor, 2017). For example, Dietrich, Enos and Sen (forthcoming) shows that the emotional pitch of audio in Justice’s arguments at the U.S. Supreme Court predicts their eventual votes. Similarly, Knox and Lucas demonstrate emotional variation in Supreme Court deliberations (Knox and Spirling, 2018). While this moves beyond the scope of machine readable text and into analysis of the audio itself, it is an important component of speeches that can be useful for scholarly use. However, the combination of multimodal channels adds non-trivial layers of complexity onto issues of document pre-processing, including aligning text with audiovisual data.

A second type of textual data growing in usage is social media data, such as Twitter data or Facebook data. These data are, in some ways, a variant on speech data: they are usually the result of an individual decision to communicate a particular thing (be it an idea, sentiment, fact, etc.) to some social audience. Like speech data, these data allow researchers to learn from the communication patterns of a massive group of people, as well as political behavior of individuals (Barceló and Labzina, 2018; Windsor, 2018). However, these data have unique characteristics. The most important quality is that they tend to be prolific over time, and cover an extremely wide array of individuals. Further, social media data provides a window into otherwise inaccessible actors: non-elite populations who normally do not communicate in ways that scholars can access. This data has allowed us to test broader theories of politics, such as inter-generational transmission of political beliefs and values (Jennings, Stoker and Bowers, 2009), as well as to test the methods that strategic governments use to disseminate selective information to populations (King, Pan and Roberts, 2017). Overall, this source of data can allow for a finer-grained measurement of non-elite politics—which, in the context of diplomacy, is useful for understanding how diplomatic information disseminates within and throughout non-elite audiences.

News reports have been used largely with application to event data (Raleigh et al., 2010; Schrodt
and Gerner, 1994). More broadly, they are a record for measuring events; from conflicts, to protests, to diplomatic summits (Bechtel and Schneider, 2010), news provides a detailed mine of data for researchers. In the example of Bechtel and Schneider (2010), for example, they use a content analysis of news reports to extract whether a summit covered European defense policy, and whether these reports were “good” news. They use these to measure whether good or bad news influence stock in European defense companies. Much like social media data, this type of analysis allows for a content-driven snapshot of communication at a fine-grained unit of analysis: in this case, for example, it allows for empirical leverage over the content of day-to-day communications of particular media outlets. This type of data, along with social media data, are useful specifically for those theories involving spatial positions expressed in political communication: in this case, for example, theories of whether relatively good or bad news from media outlets affect stock prices.

Finally, the broadest category of text would be formal, institutional documents—from media reports to legal reports, and even diplomatic texts (such as cables) themselves. These types of documents tend to be relatively structured, and have seen applications ranging from event data gathering (Raleigh et al., 2010; Schrodt and Gerner, 1994) to using Human Rights reports to quantify human rights violations (Greene, Park and Colaresi, forthcoming). These institutional documents provide several opportunities. First, from a data collection standpoint they provide a systematic, streamlined repository of information: institutional documents tend to be relatively well structured, and as a result can make parsing and organizing data much easier. Second, these documents provide a window into the outcomes of institutional processes. Perhaps the most well-known example of this would be the extensive work that has been done on measuring party positions via the Comparative Manifesto Project text (Laver and Garry, 2000; Laver, Benoit and Garry, 2003; Slapin and Proksch, 2008a; Lowe et al., 2011). These texts provide a systematic view of how political parties aggregate their public communications, and can be especially useful for testing theories of aggregating preferences through institutions—such as by comparing positions expressed in institutions to those expressed by individual actors within an institution.

**Concepts and assumptions**

The use of text documents as empirical information depends on several assumptions, depending on the nature of the document, the research question, and the measurement strategy. Thus, while we outline what we believe are some of the most common, and most important issues below, each research question and data source comes with its own set of assumptions and issues, and thus should be given its due diligence.

**Document level of analysis**

A common use of these data sources is to use text to measure *actor positions* relative to one another. Applications to Political Science have largely focused on political parties (Slapin and Proksch, 2008a; Monroe, Colaresi and Quinn, 2008; Laver, Benoit and Garry, 2003), drawing on either party manifestos or legislative speeches as data. These works contribute to broader efforts
over the past several decades to quantify new sources of data in order to understand political actors; these efforts are most largely associated with the surging use of item-response models and other measurement models to convert large amounts of noisy data into an averaged, systematic view of the actors involved (Linzer and Staton, 2015; Bailey, Strezhnev and Voeten, 2015; Martin and Quinn, 2002; Voeten, 2000; Poole and Rosenthal, 1997, 1985). An ongoing debate exists, however, in how to extract these measurements while accounting for bias (e.g. Benoit, Laver and Mikhaylov (2009a)). Beyond statistical bias, as a matter of theory it becomes useful to evaluate texts in the context of their data generation: for example, strategically generated political documents may be more or less informative (as in, systematically correlated with outcomes or concepts of interest) in some contexts over others. The question of understanding the conditions under which particular diplomatic texts are informative, and the nature of mapping this text to the concepts in question, provides a useful opportunity for theorists of diplomacy to work with this type of data.

It is important, however, for researchers to appreciate the crucial assumption in all of these studies: namely, that these texts contain systematic information about the underlying patterns of interests and preferences. For example, the use of political speeches to measure the interests of political actors directly implies that these political actors are actually communicating their interests—in some way—through these speeches in a systematic, informative way. In summary, a critical assumption in using a body of text to measure information is the informativeness of that text. As Benoit et al., demonstrate, we can extract policy positions from textual data to estimate policy positions from corpora such as the Comparative Manifesto Project (Benoit, Laver and Mikhaylov, 2009b).

A second use of these texts is to compare two actors (or two expressions of text for the same actor) not on their positions, but on the characteristics of that position. For example, McManus’s (2014) use of text to measure resolve assumes that not only do the texts contain information about resolve, but also about the ordinality of this resolve. A similar assumption is true of using party manifestos: not only does this work assume manifestos systematically communicate interests, but that they also communicate the salience of a particular interest. For example, in Slapin and Proksch’s (2008b) work to scale German parties, their measurements show that the “[Free Democrats Party (FDP)] tends to be slightly to the right of the [Christian Democrats (CDU-CSU)] up until 2005, when it moves to the center”. This implies that their measurements capture not only positions (left versus right), but also the degree to which both parties are further right and/or further left from one another. Overall, a secondary assumption is that text is not only informative, but that it scales in a lower dimensional space to place actors ordinally relative to one another (Lowe, N.d.).

Finally, in addition to informativeness and ordinality, scaling multiple actors at once implies that text is comparable in its data generation process. For example, in the study of international diplomacy, comparing two speeches delivered on the floor of the UN implies that both speakers are communicating information on a comparable scale. Work by Baturo et al., has demonstrated the utility of UN public debate for estimating countries’ agendas (Baturo, Dasandi and Mikhaylov, 2017). This assumption is relatively simple to make in domestic legislatures, where the majority of Political Science work on text and speech has taken place: for example, in the United States, it is reasonable to assume that both Democrats and Republicans can be mapped to the same left-right political dimension. By scaling multiple actors to the same dimension(s), a researcher assumes comparability between the actors with respect to the text—be it a document, speech, or media
In addition to the assumptions and concepts that go into the use of particular documents, it is important for researchers to take seriously the assumptions that come with various \textit{corpora}. Beyond questions of selection bias, which we discuss later, researchers must take care to understand whether the corpus used for analysis actually provides the information needed. For example, the text of a UN resolution–which can be the result of negotiations, vote buying, and more between countries of varying interest and power–is very different than the text of a UN Secretary General Report, composed with slightly greater independence from those individual actors. Similarly, texts of speeches delivered to one organization might be different than speeches delivered to another organization, depending on a variety of factors in both organizations. The \textit{data generating process} of how a corpus is generated is an important part of selecting and using text to test theories.

\textbf{The flexibility-quantity tradeoff}

Qualitative methods deserve a unique emphasis with regard to text data. Language is, by its very nature, \textit{subjective}: while computational linguistics has made substantial progress in parsing the structure and meaning of sentences, the human eye remains the most faithful reading of text, since text is written and/or spoken for human consumption and interpretation. As Laura Nelson noted, computers perform well with sorting and optimizing; humans, on the other hand, excel at sense-making and interpretation, and computational discourse analysis should exploit and leverage these relative strengths to analyze large-N corpora (Nelson, 2017). One potential solution to address problems of corpus generation and selection bias is that social scientists and computer scientists collaborate on research projects. Again, this approach would leverage the relative strengths from each discipline: social scientists are trained to ask questions about social and political relationships, but often lack technical skills for generating, cleaning, and coding textual data. On the other hand, computer scientists excel at automating the process of generating text data, but are not trained to contextualize the data within social science frameworks.\footnote{An interesting concern, however, should be of note here: with the increasing prevalence of automated parsing, it remains a possibility that over time, speech and written text will become more and more tailored to these methods of interpretation. As an example, consider whether listing specific skills on a resume or cover letter becomes more important given the growing popularity of automated sorting algorithms in corporate hiring processes.}

\textbf{A utilitarian view of text-as-data}

A utilitarian view of text-as-data–that is, focusing on text as a \textit{tool} to achieve a particular goal most efficiently–can help circumvent the problems that come with using text to measure a theoretically specific concept. This is most easily illustrated by the use of text as simply a high-dimensional source of data–with unknown ratios of signal and noise–as a tool to build increasingly accurate forecasts or
simulations. In these cases, while theory is still necessary to broadly specify the selection of features to include in a model (e.g. a theory linking protests to civil war to justify using event data from news reports), the utilitarian view of text as data can relax the potentially stringent induction-deduction loops that come from more traditionally focused research designs, without necessarily sacrificing the capacity to learn from the data (Colaresi and Mahmood, 2017). In this view of text as data, scholars can potentially gain high leverage over a well-specified task (for example, predicting events) by allowing flexible models to optimize over large feature sets (such as text) without necessarily having each individual feature be theoretically vetted a priori. Note, however, that even under this use of text-as-data, there are still important questions to consider in the actual mechanics of using the text, including the implications of different pre-processing procedures prior to model estimation (Denny and Spirling, 2018).

Challenges and limitations of the data sources

There are a number of challenges when using text-as-data. First, there may be biases in data collection that limit the generality of information gathered from the sample to the population. Next, issues related to language and context can impact on how researchers interpret output generated from texts. Finally, computational techniques used to analyze text-as-data can impose model structures on the data that affect the substantive inferences that researchers draw.

Computational text analysis deserves thoughtful consideration of selection bias: As Geddes notes, the cases you choose affect the answers you get (Geddes, 1990). Much of the existing scholarship within Political Science using quantitative text-as-data methods relies on corpora from Western sources, such as legislatures and courts, and new media sources like Twitter. These corpora require less pre-processing than do corpora from less developed nations, including optical character recognition (OCR), translation, and text cleaning with regular expressions to remove erroneous elements from the documents (Denny and Spirling, 2018). While somewhat tedious and unglamorous, the significance of such pre-processing steps cannot be overstated: it is imperative that scholars examine political processes in their original languages, or reliably translated versions, whenever possible.

Corpus selection bias generally arise from information deficiencies. Information deficiencies arise when documents are unavailable because of poor archiving processes, incomplete records-keeping due to conflict or inclement climate conditions, or because the curators wish to keep the documents private - as in authoritarian regimes or terrorist groups. One well-known example are event data gathered from media outlet reports. Media reports systematically under report events of interest to political scientists due to lack of opportunity (no reporter near event) or willingness (perceived lack of consumer interest) (Cook et al., 2017; Bagozzi et al., 2018). Likewise, reports from government agencies and third-party observers may reflect the strategic interests of these groups (Hill, Moore and Mukherjee, 2013; Nieman and Ring, 2015). Notably, these problems stem from the reports themselves, meaning that neither human- nor machine-codings provide a panacea to possible selection or reporting bias.

Moreover, difficulties with collecting random samples are not limited to media reports. There is
nothing about Twitter’s data collection or scrapping rules, for example, that ensures a random sample. Over-reliance on new media may introduce other types of bias as well. For instance, Twitter data may not produce reliable results about global phenomena, as a non-trivial proportion of tweets originate from users in countries outside the region of interest, as happened in the Arab Spring and in Iran’s Green Revolution. Twitter scraping algorithms also do not produce random samples, and are plagued by the intrusion of bots that introduce noise.

Text-as-data also face limitations due to document processing, impacting a researcher’s substantive understanding of a corpus. Document processing problems arise from a lack of parallel corpora from which to generate semantic spaces or reliable translations, and a lack of partnership with primary language speakers and scholars who could provide useful insight not only into the language, but the political context in which it occurs. Many of the computational programs used to analyze syntactic properties of language work predominantly on English-language corpora, although advances in computational linguistics and computer science will likely improve this scenario in the near-term. Scholars may opt to use translations, although this will likely produce errors by not preserving the original structure of the source language. For example, passive voice varies widely between languages, as does sentence boundary disambiguation. What constitutes a sentence is not a constant across languages. Many other processes such as Latent Dirichlet Allocation are language-agnostic, and can accommodate multilingual corpora (Lucas et al., 2015). For example, the use of n-grams must account for varying language structures when attempting to put words in context.

Semantic and syntactic differences between news reports can also introduce bias; most event data is generated from wire service reports that produce highly formulaic and SVO (subject-verb-object) headlines and article structures. Local news sources, however, may be more ornate and descriptive, styming the NLP parsers accustomed to the SVO format. ICEWS notably missed the onset of the Arab Spring, in part due to the reliance on global news media rather than local sources. Furthermore, international actors and news sources can describe the same event in different ways, signifying that they perceive the ontological nature of the event differently. As such, the various codes assigned to the event (i.e., CAMEO or Goldstein) may vary depending to who is describing it. At present, different cultural and linguistic interpretations of events are not accounted for in generating event codes, but rather coded based on Western interpretation of events.

Finally, models that require a priori specifications impose a specific structure on the data. Topics models, for example, may find patterns in the data that are substantively non-meaningful. While a topic consisting of words of set including ‘investment’, ‘trade’, and ‘tariff’ from a corpus of treaty documents may be easy for a research to classify as ‘economic treaties’, other identified topics may not be so easy. Not all topics may represent meaningful groupings, no matter our tendency to find patterns in the noise. This is especially problematic if, as is often the case in political science, we do not have fully know what we are looking for within the data. Drawing inferences from textual data requires a degree of care and humility on the part of the researcher. As Denny and Spirling (2018, 169) note, the “forking paths” arising from early decisions can result in very different inferences, especially for unsupervised techniques. This level of care implies that the extensive content and explicit descriptions of the methods, so that others can replicate results, are of even greater importance than with other types of data analysis.
Example data sources

The growing text-as-data field has a new repository for software, conference archiving, scholar networking, and data citations: https://QuantText.com This domain provides a centralized location for text-as-data scholarship.


References


