Review Article

Advancing the Field of Writing Analytics: Lessons from "Text-as-Data" in the Social Sciences

Ian G. Anson, *University of Maryland, Baltimore County*

Abstract

The field of writing analytics is currently in a state of growth, redefinition, and refinement. In this essay, I review the trajectory of a related subfield, text-as-data in political science, as a lens through which to understand the present and future of writing analytics. I first describe how text-as-data has evolved over several eras, before transitioning to a review of some of the most exciting contemporary areas of political text-as-data. I then draw parallels between these developments and the work being done by the community of writing analytics scholars across the globe. I conclude by proposing several considerations for current practitioners seeking to emulate the "success story" of political text-as-data, including investment in the collection of new, high-quality corpora, development of shareable and open-source methodological tools for text analysis, and the strengthening of a community of scholarship.

Keywords: **corpus construction, text-as-data, writing analytics**

1.0 Introduction

The field of writing analytics is currently in a state of growth, redefinition, and refinement. Since the publication of the first issue of *The Journal of Writing Analytics* (*Analytics*), we have witnessed the emergence of a community of scholars characterized by commitments to multidisciplinarity, reflection, and evidence-based propositions (Lang et al., 2019). In this review article, I briefly outline the disciplinary history of text-as-data in political science, a field with similar, but not identical, foundational principles. My goal is to examine how this burgeoning

area of research, characterized by its use of quantitative tools for the analysis of political texts, can inform the present and future development of writing analytics.

While social scientists have engaged in the empirical study of texts over many decades, recent technological innovations have reinvigorated such efforts. In political science, scholars have borrowed tools from computer science and computational linguistics, applying them to the large-scale examination of texts using techniques like sentiment analysis, document classification, causal analysis, and textual network analysis. These innovations have led to a diverse set of discoveries in fields such as presidential behavior, congressional credit-claiming, political behavior, political media, and public opinion. Together, they have led to novel theorizing and major methodological advances.

Text-as-data's growth over the past two decades can serve as a useful lens to understand the state of our own discipline. In this article, I identify three principal reasons why political text-as-data has emerged as an invigorated subfield in political science—despite the traditional dominance of other empirical approaches (Dahl, 1961). First, as computational text analysis became more accessible and less costly, political scientists invested in the creation, curation, and dissemination of new large-scale political corpora. These corpora have permitted researchers and research teams to replicate earlier findings, build new models of unexplored features of text, and most critically, to *apply* these text-based discoveries as independent variables in new studies of diverse political phenomena.

The second reason for political text-as-data's proliferation has been the rise of open-source programming tools created by scholars for public use. While in earlier eras, scholars might have built code to analyze their own proprietary corpus, political scientists are now building open-source software packages in languages like R and Python which can be flexibly and robustly applied to a wide variety of texts. By hosting these programs on stable, vetted repositories like PyPI, the R-CRAN network, and GitHub, the user base has benefitted over the long term.

Finally, political text-as-data practitioners have invested in a community of scholarship—something that writing analytics has also achieved through the ongoing efforts of conference organizers both in Europe and the United States (Moxley et al., 2019). In the concluding section, I discuss ways in which this international network of text analytics scholars can benefit from the aforementioned strategies pioneered in political text-as-data.

These recommendations boil down to a call for increased collaboration. To see writing analytics grow in the 2020s we should find new ways to share our data, resources, tools, ideas, and strategies. I conclude with two principal recommendations for writing analytics:

- Develop and host new corpora for free, open-source distribution to the community of text-as-data practitioners.
- Curate a repository of robust, well-commented code and/or packages for popular programming languages, that help practitioners understand extant approaches to available corpora.

These steps will require cooperative efforts, spearheaded and organized by the field's pioneers and newcomers alike. In the next sections, I introduce the scholarship of text-as-data in political science, before describing how scholars in that field have organized their own collaborative effort to share tools, resources, and data.

2.0 The Nascence of Political Text-as-Data

Political scientists have been interested in text analysis for at least a century. In the earliest era of political science, many scholars occupied themselves with close qualitative reading of political primary sources. Notable early political scientists like Woodrow Wilson wrote theory-driven accounts of political history in this manner, explaining the rise and fall of political movements, the motivations of political actors, and the fundamental question of "who gets what" in modern democracies (Schattschneider, 1975). One salient example is the writing of Charles Beard (1913), whose scrutiny of the Constitution and the Federalist Papers would challenge historical accounts by situating the Framers as self-interested commercial actors.

While these early qualitative approaches reflected the dominant methodological techniques of the discipline's earliest years, soon the field would accommodate new methods from sociologists and early social psychologists. The behavioral turn in political science germinated in the 1940s and grew to become a dominant paradigm from the 1950s to the 1960s (Dahl, 1961; Easton, 1969; Farr & Seidelman, 1993). Owing partially to the influence of Max Weber and European sociologists, the behavioral approach would come to be characterized by a system of naive empiricism and data-driven claims. David Easton, a key figure in this revolution, explained in the early 1950s that behavioralism rests upon several "intellectual foundation stones," among them the examination of regularities, a commitment to verification, a reliance upon specialized techniques, the use of quantification, the practice of "value-neutrality," systematization, and an overarching adherence to the practices of "pure science" (1953).

The behavioral revolution changed the way political scientists studied texts. Adherents demanded a rigorous, non-ideological, and replicable methodology for making any scientific claims—especially because the term *social science* was beginning to sound too much like *socialist science* to many conservative politicians (Seidelman, 2015). Some political scientists began employing more structured qualitative content analyses in order to study texts. To do so, they adopted methods from sociology and the nascent study of mass communication. Early work examined the content of presidential speeches and Supreme Court decisions, among other subjects (e.g., Berelson, 1952; Prothro, 1956; Schubert, 1958; White, 1956).

Nevertheless, these qualitative studies of text-as-data were partially overshadowed by the development of modern surveys. The "Michigan School" of behavioral research, also emphasizing Easton's value-neutrality and quantification, spearheaded a move towards survey research that had far-reaching consequences. Presently, the field of American politics still privileges survey methods, as we can see from the hundreds of recent books and articles whose evidence derives from the longstanding American National Election Studies (http://www.electionstudies.org). As "American-style" political science grew by leaps and

bounds over the decades, spreading across the world and becoming far more methodologically and paradigmatically diverse (e.g., Almond, 1990), the survey has retained its hegemony in spite of the rise of other peacefully coexisting approaches.

3.0 Towards "Political Text-as-Data"

Since the 2000s, however, political science has witnessed a return to text analysis in an effort to answer new questions about elite and mass politics. In pursuit of the same behavioral principles of quantification and value-neutrality, this next wave of text analytics turned to quantitative, rather than qualitative, tools.[1] The rise of computational linguistics has allowed practitioners to move away from labor-intensive qualitative content analysis to fully- or partially-automated procedures for extracting meaning from natural language.

Of course, readers of *Analytics* are likely well aware that some of these tools were developed long ago. Some pioneering attempts at quantitative text analysis can even be glimpsed in the pages of political science's flagship journals as early as the 1970s (e.g., Cary, 1977), though these efforts did not catch on. In a meta-analysis, Popping (2000) shows that political science actually saw a *decrease* in the already scant use of quantitative text analytical approaches from 1970-1986.

Instead, painstaking, large-scale human-coded content analyses like the Comparative Manifestos Project (CMP) were launched in the 1980s and 1990s (Budge et al., 1987; see also Tetlock, 1981 for a separate example of qualitative content classification). The CMP (now simply called the Manifesto Project) is a multinational study of the policy positions of political parties, based on text analysis of those parties' official platforms (http://manifesto-project.wzb.eu). It would not be until the 2000s that political scientists, including the members of the CMP, began employing large-scale automated text analysis more regularly. Some pioneering efforts by Laver, Benoit, and Garry (2003; see also Laver & Garry, 2000) and King and Lowe (2003) appeared in the early 2000s, forerunning a wave of quantitative text analytical research on party platforms, elite communication, constitutions, and court decisions.

Modern text-as-data now encompasses a variety of topics and approaches, among them analyses of "core" institutional texts, the behavior and communication of political elites, political media, and even public opinion and mass politics (the lattermost historically serving as the principal redoubt for survey scientists). In the next sections, I review each of these topics in turn, briefly detailing the major works, methodologies, and approaches that have contributed to the reinvigoration of political text-as-data.

---

[1] This is not to say that political text-as-data has unilaterally *succeeded* in developing a "value-neutral" social science. Many critiques of modern "big data" processes have pointed to the possibility that human biases will be compounded, rather than mitigated, by algorithms and unsupervised learning (e.g., Crawford & Calo, 2016). While many political text-as-data practitioners likely take the premise of value-neutral social science as a grounding principle, critiques of value-neutrality in quantitative political science research have become highly salient in recent years (e.g., Seidelman, 2015).

3.1 Studies of Political Institutions

While texts like the U.S. Constitution were once the principal data source for early political scientists' literary examinations, "core" institutional texts are now an exciting resource for computational analysis. Supreme Court opinions, laws, constitutional texts, bureaucratic rules, and policies are now studied using tools like unsupervised classification and topic modeling. In an early effort, Evans et al. (2007) classified the ideological positioning of advocacy briefs (*amici curiae*) submitted to the Supreme Court, showing that supervised machine learning could be used to reliably investigate Court decision-making.

Additional language features, including lexical complexity and tone, have been examined by scholars of political institutions in recent years (e.g., Black & Owens, 2011; Black et al., 2011; Owens & Wedeking, 2011). These studies explore judicial decisions and other political texts with high levels of descriptive detail, tapping emotional dynamics and linguistic style to inform hypotheses about Court precedent and the long-term impact of judicial decisions.

Political parties, their ideological positions, and elite partisan discourse represent promising directions for study as well. As described above, some of the earliest text-as-data forays in political science addressed questions in this subfield (Benoit et al., 2009; Laver et al., 2003; Laver & Garry, 2000). More recent research has built on this foundation in a variety of ways. Spirling (2016) and coauthors (e.g., Denny & Spirling, 2018; Peterson & Spirling, 2018) have focused on the British Parliament, examining a variety of features of parliamentary language across history. These investigations have taught us a great deal about how legislators compete and organize along ideological fault lines. The development of specialized tools for this substantive area of research has also allowed Spirling (2012) to examine other topics of relevance to political historians, such as treaties made between the United States and American Indians across history.

So too have scholars such as Diermeier et al. (2012) been able to use text-mining tools to predict ideological positioning in the U.S. Congress. Drawing on these approaches, students of federalism and state politics have also made important contributions. Text similarity algorithms have helped scholars determine when and where policy ideas are passed from one state to the next, and from one bill to another (Casas et al., 2020; Wilkerson et al., 2015). The ability to trace policy diffusion, an important principle in the study of American federalism, is yet another affordance of text-mining techniques. Some of the most recent research in this area has sought to examine the growth of party polarization: Goet (2019), for example, recently applied supervised classification techniques to model legislative polarization in the UK Parliament over the past 200 years.

3.2 Studies of Elite Discourse

Moving beyond party platforms, institutional frameworks, and other formal political texts, political scientists have also become interested in the features of elites' communication efforts. These texts might include committee transcripts, candidate speeches, lobbying documents, political advertisements, press releases, and other attempts to influence peers and/or the public.

Text-as-data approaches in this field help researchers to develop more granular and robust knowledge when it comes to topics such as legislative influence and the role of lobbying in congressional representation.

Perhaps the most well-known of these recent efforts have been conducted by Justin Grimmer (2013) and coauthors (e.g., Grimmer et al., 2012; Grimmer et al., 2014). Several of these investigations rely on a large repository of press releases made by incumbent politicians in the U.S. Senate, a dataset that can be mined for sentiment, topic classification, and other key features. The authors of the aforementioned studies have used this repository to examine legislator credit-claiming and blame-giving, along with other aspects of "representational style," including the willingness to emphasize "pork" (appropriations) over substantive policy.

Many other forms of elite political communication have also been recently examined by text-as-data scholars. From State of the Union addresses to floor debate in Congress, scholars have been able to use text repositories to more precisely examine patterns in elite discourse (Benoit et al., 2018; Gentzkow et al., 2016; Herzog & Benoit, 2015; Quinn et al., 2010). These studies are beginning to encompass politics outside of the US and the UK as well: Bustikova et al. (2020), for example, recently showed that latent Dirichlet allocation (LDA) can help identify when and why Slovakian parties respond to their political rivals.

3.3 Studies of Media and Politics

Of course, in an era of populist communication styles, any discussion of elite discourse inevitably invokes the subject of social media. Tweets, Facebook messages, online advertisements, Instagram posts, and a bevy of other social media sources represent promising venues for the study of elite and mass discourse. Beginning with the work of Bollen et al.(2011), political scientists have become enamored with the use of text-as-data tools to study social media, its causes, and its effects (e.g., Barberá et al., 2019; King et al., 2017;). A notable contribution in this field is that of Barberá et al. (2015), who use Twitter to partially debunk popular assumptions about online selective exposure (or the "echo chamber" effect). And Beauchamp (2017) cleverly uses location-based Twitter data, paired with topic modeling, to make electoral predictions that are more accurate than state-level presidential election polls alone. Jamal et al. (2015) study anti-Americanism on Twitter in Arabic-speaking countries, highlighting the ways in which Twitter scraping, sentiment analysis, and topic modeling can be applied to international political questions as well as domestic ones.

Social media may be a hot topic for many scholars of text-as-data, but political scientists (especially those who specialize in communication studies) have also made great progress over the last decade in studying the news. Agenda-setting, partisan slant, tone, and other properties of news content have been explored on a massive scale thanks to new text-as-data techniques (e.g., Boydstun, 2013; Gentzkow & Shapiro, 2010; Young & Soroka, 2012). Scholars in this realm have paid special attention to the subject of economic news, in an effort to better understand the foundational theory of "economic voting" in American elections (e.g., Boydstun et al., 2018; Soroka et al., 2018).

3.4 Public Opinion and Mass Politics

As mentioned above, public opinion represents an area of political science that has been historically dominated by survey methods. Nevertheless, exciting text-as-data research in the study of public attitudes has reinvigorated the field. Some recent work has used structural topic models to better understand dynamics in public opinion, drawing on the affordances of surveys' open-ended responses (e.g., Roberts et al., 2014). By finding more sophisticated ways to process and classify the natural language captured in many historical surveys, these studies have advanced our understanding of topics like ideology and party attachment.

Very recent examples of text-as-data in public opinion reveal the great promise of this approach. To better understand political sophistication, a key concept in the study of mass and elite attitudes, Benoit et al. (2019) paired random forest algorithms with a hand-coded set of lexical tokens. This large-scale analysis produced a method for judging sophistication in a wide variety of political texts. And Emma Rodman (2020) has recently advanced our understanding of ideology and political sophistication by focusing on the underlying *meaning* of political concepts as discussed in text over time. This semantic analysis, which relies on word vectors[2], is yet another way in which text-as-data is refining the body of knowledge in American politics.

3.5 Text Analytical Tools in the Pages of *Political Analysis*

Of course, none of these studies would be possible without the work of political methodologists, who have developed new and more sophisticated tools for studying text. In political science's flagship methods journal, *Political Analysis*, the most recent issues have been replete with such papers. Building on earlier methods, such as the Wordscores approach of Will Lowe (2008) and the Bayesian hierarchical topic models of Grimmer (2010), these recent methodological advances demonstrate the diversity of approaches inherent to the political text-as-data community.

Recent work in *Political Analysis* has introduced advanced techniques for classification, linguistic feature extraction, sentiment analysis, and text network analysis, especially where corpora possess technical challenges (Grimmer, 2013). In one notable example related to document classification, D'Orazio et al. (2014) introduce a support vector machine approach to engage in fully-automated classification of very sparse documents. Other approaches have echoed the recent focus in political science on causal analysis (e.g., Egami et al., 2018; see Keith et al., 2020 for a review). By pairing text analytical tools with the causal inference approach

[2] Word vectorization is a process by which natural language is converted into numerical vectors which can be mapped spatially. Thus, text similarity and other features can be assessed using numeric computations. An example of this approach is cosine similarity, which evaluates the similarity of texts by comparing the coordinates of word vectors.

pioneered by scholars like Rubin (2006) and Pearl (2009), these authors have opened the door for hybrid studies that make causal claims using observational text data.

Other innovative scholars in the discipline are beginning to reimagine how texts might be used for empirical applications. Harris (2015), for example, studies how text-as-data approaches can be leveraged to accurately classify names according to demographic characteristics, aiding in studies of racial and ethnic politics and gender and women's studies. And Proksch et al. (2019) have recently explored ways to glean meaning from spoken recordings, using automated speech recognition software to convert speeches and other recordings into tractable text-as-data structures. These techniques showcase the ways in which political scientists, working at the intersection of methods and substantive knowledge, are expanding the boundaries of the possible in the quantitative analysis of text.

4.0 Political Text-as-Data: Roots of Success

Despite having only partially reviewed the recent contributions of political text-as-data in the preceding sections, the field's ongoing success seems evident nonetheless. Scholars of text-as-data now place their work in the top outlets in the discipline with regularity, and text-as-data's methodological innovations are being used with increasing frequency outside of political science. More important still, the substantive contributions to our understanding of real political phenomena are plain to see in the pages of recent books and articles. What, then, explains the recent explosion—much akin to the "uncorked champagne bottle" described by Moxley et al. (2017) in this journal's inaugural issue—of political text-as-data?

4.1 Shared Corpora

Text-as-data has been able to grow much faster than it would have otherwise due to its commitment to developing and sharing text data resources with other scholars. One pioneer in this approach has been Stuart Soroka, whose sentiment analysis allows for easy and widely-applicable analysis of sentiment in political texts (http://www.lexicoder.com). This corpus of tagged text can be applied to other corpora in order to produce vetted measures of sentiment. Another example is Amber Boydstun's Media Frames Corpus (https://www.aclweb.org/anthology/P15-2072/), which has allowed scholars of political communication to obtain a valuable resource for the study of framing and news slant.

Justin Grimmer's repository of Senate press releases is an especially valuable resource for the study of elite discourse. This corpus, which contains all official Senate press releases from the 109th through the 111th Congress, allows scholars to model legislative behavior and apply those insights to key outcomes of the legislative process. When it comes to presidential communication, the Miller Center at the University of Virginia has compiled a crucial resource: thousands of freely available presidential speeches dating back to George Washington. This repository is also continuously updated, meaning that the repository contains very recent presidential communications as well as historical ones.

More generally, shared corpora reflect political science's turn towards open science. Open science is characterized by a commitment to increasing the openness, integrity, and reproducibility of research, as emphasized by the mission statement of the Center for Open Science (https://osf.io/x2w9h/). Of special importance to text-as-data practitioners is the ability to replicate published findings by obtaining access to a paper's underlying code, data, and software. As text analytics rely on "big data" approaches and the use of sophisticated algorithms, this openness imperative requires that scholars host their data sources and code on reliable platforms like GitHub (http://github.com). When code is written using free open-source software like R and Python, the barriers to replicability are lowered even further.

Openness can also be facilitated by the curation and management of replication materials by journals. Special data repositories have been created for this purpose, such as the University of Michigan's ICPSR (Inter-University Consortium for Political Science Research) dataverse (http://icpsr.umich.edu) or the Odum Institute Data Archive at the University of North Carolina (http://odum.unc.edu). Many political science journals have recently written statements supporting open science initiatives and have partnered with high-quality data repositories to support that mission. Upon an article's acceptance in a journal, replication materials are checked by journal staff for accuracy and are then stored in a data archive for public use.

As practitioners, the ability to access the data required to replicate existing findings also yields new avenues for research. Repositories host large-scale text databases that can be used for new, unanticipated applications. Take, for instance, Grimmer's (2013) repository of Senate press releases. This set of texts can be used by scholars studying topics as diverse as social media, economic voting, minority representation, and lobbying as a context for comparison, a source of "independent variables," or as a subject of study. In earlier years, corpora like this one would only be available by request from authors—and requests might not always be fulfilled.

4.2 Shared Tools

The rise of open science has also led to an emphasis on the development of shared tools for analyzing text data. Thankfully, departments and programs are beginning to formally recognize the contributions of these tools as they would other forms of publication or creative achievement, leading to still more software development in the field. Examples include Soroka and Young's Lexicoder sentiment software , the "quanteda" package developed by Benoit et al. (2018), the R "stm" package developed by Roberts and colleagues (http://structuraltopicmodel.com), Pablo Barberá's "streamR" package for importing Twitter API data (https://cran.r-project.org/web/packages/streamR/streamR.pdf), and Barberá and colleagues' Rfacebook package for acquiring large amounts of Facebook data through the Facebook API (https://github.com/pablobarbera/Rfacebook).

Other examples from political science abound, and are most often written in the R programming language and hosted on the R-CRAN network of vetted packages (http://cran.r-project.org). While this approach can place a steep learning curve on the use of text-as-data tools for non-R users, the R language continues to evolve into a more user-friendly platform. This

initiative has been spearheaded by Hadley Wickham and proponents of the "tidyverse" approach (http://tidyverse.org). In particular, the use of libraries associated with the "tidytext" approach to text-as-data (Silge & Robinson, 2016; http://www.tidytextmining.com) has increased the accessibility of text-as-data tools. As the scholarly community grows, and as more practitioners continue to replicate and build upon existing work, the tools have also been refined through user-generated troubleshooting and bug reports.

5.0 The Writing Analytics Community in the 2020s

This incomplete review of political text-as-data is intended to illuminate the "secrets" of a successful disciplinary subfield. While some of my points echo those made in earlier pages of *Analytics*, my hope is that the story of text-as-data will advance our own conversations in several ways.

First, I hope that it showcases areas in which text-as-data's approach has differed from that of writing analytics. Our discipline has a far different substantive orientation, and its guiding principles reflect some of those commitments—most notably a commitment to students and instruction. This has led to a commitment to reflection and ethical philosophy that uniquely positions writing analytics relative to subfields like learning analytics (Lang et al., 2019; Moxley et al., 2017) and political text-as-data. The richness of writing analytics comes in part thanks to its deep self-awareness in this regard.

Writing analytics is also far more interdisciplinary and multidisciplinary than political text-as-data. This is despite the fact that text-as-data is also positioned at disciplinary crossroads, borrowing many techniques and approaches from statistics, computer science, communication studies, and sociology. The remarkable diversity of writing studies—from a disciplinary, substantive, and goal-orientation perspective—is both a strength and a challenge in comparison. Political text-as-data has been able to successfully organize in part because its practitioners mostly identify as political scientists. Conferences like the annual Text as Data Association (TaDA) meeting (http://textasdata.github.io/) serve to organize text-as-data practitioners, whose presence on political science mailing lists and in disciplinary professional organizations facilitate calls for papers.

In comparison, writing analytics has sought to organize through its journal, *Analytics*, and especially through the International Conference on Writing Analytics. This event has served as a way to grow the field through its inclusive call for presentations and its array of workshops and pre-conference activities. *Analytics* has worked over the past several issues to define, taxonomize, and expand writing analytics, carefully spelling out the field's commitments and core principles.

Yet, while this work continues apace, the 2020s offer writing analytics an opportunity to develop centralized resources for new scholarship in the spirit of recent text-as-data initiatives. During the earliest years of *Analytics*, many published studies were based on analyses of the University of South Florida's My Reviewers corpus (Branham et al., 2015). In the next phase of writing analytics scholarship, new corpora should be developed and shared with the scholarly

community, along with the code necessary for their analysis. Centralized repositories of corpora could be hosted by *Analytics* and shared with interested scholars.

This open science approach can also be encouraged by creative amendments to *Analytics* guidelines. For instance, many political science journals have recently moved to accept submissions in the form of pre-analysis plans (see, for example, the guidelines of the *Journal of Experimental Political Science*: https://www.cambridge.org/core/journals/journal-of-experimental-political-science/information/faqs-for-registered-reports). Pre-registered analysis plans of new corpora (or corpora in development) would allow practitioners to propose novel hypotheses and methods while receiving useful early feedback about study design and framing. Simply put, by organizing our efforts around shared data resources and pre-analysis feedback, practitioners would be poised to contribute more and better findings to our collective base of knowledge.

Finally, pre-conference activities such as "hackathons" and tutorials would be immensely aided by a centralized repository of writing analytics data. This would allow practitioners to try out advanced text analytical tools, learning new techniques for their own substantive research in the process. The next phase of writing analytics research is poised to produce a large volume of rigorous empirical studies on writing and learning, much as the past decade in political science has witnessed the rise of text-as-data. With new, creative ways to organize and share resources, the work ahead will be far easier and more rewarding.

Author Biography

Ian G. Anson is assistant professor in the Department of Political Science at UMBC. Prof. Anson's research interests encompass a diverse array of topics, from the scholarship of teaching and learning to subjects in public opinion and political psychology including partisan motivated reasoning, misinformation, political knowledge, and responses to elite cues. Prof. Anson received a PhD in political science and an MS in applied statistics from Indiana University in Bloomington, Indiana.

References

Almond, G. A. (1990). *A discipline divided: Schools and sects in political science*. Sage.

Barberá, P., Casas, A., Nagler, J., Egan, P. J., Bonneau, R., Jost, J. T., & Tucker, J. A. (2019). Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review*, *113*(4), 883–901.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531–1542.

Beard, C. (1913). *An economic interpretation of the Constitution of the United States*. Macmillan.

Beauchamp, N. (2017). Predicting and interpolating state-level polls using Twitter textual data. *American Journal of Political Science*, *61*(2), 490–503.

Benoit, K., Laver, M., & Mikhaylov, S. (2009). Treating words as data with error: Uncertainty in text statements of policy positions. *American Journal of Political Science*, *53*(2), 495–513.

Benoit, K., Munger, K., & Spirling, A. (2019). Measuring and explaining political sophistication through textual complexity. *American Journal of Political Science*, *63*(2), 491–508.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). Quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774.

Berelson, B. (1952). *Content analysis in communication research*. Free press.

Black, R. C., & Owens, R. J. (2011). Solicitor general influence and agenda setting on the US Supreme Court. *Political Research Quarterly*, *64*(4), 765–778.

Black, R. C., Treul, S. A., Johnson, T. R., & Goldman, J. (2011). Emotions, oral arguments, and Supreme Court decision making. *The Journal of Politics*, *73*(2), 572–581.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8.

Boydstun, A. E. (2013). *Making the news: Politics, the media, and agenda setting*. University of Chicago Press.

Boydstun, A. E., Highton, B., & Linn, S. (2018). Assessing the relationship between economic news coverage and mass economic attitudes. *Political Research Quarterly*, *71*(4), 989–1000.

Branham, C., Moxley, J., & Ross, V. (2015). My Reviewers: Participatory design & crowd-sourced usability processes. In *Proceedings of the 33rd Annual International Conference on the Design of Communication* (pp. 1–6). Association for Computing Machinery.

Budge, I., Robertson, D., & Hearl, D. (Eds.). (1987). *Ideology, strategy and party change: Spatial analyses of post-war election programmes in 19 democracies*. Cambridge University Press.

Bustikova, L., Siroky, D. S., Alashri, S., & Alzahrani, S. (2020). Predicting partisan responsiveness: A probabilistic text mining time-series approach. *Political Analysis*, *28*(1), 47–64.

Cary, C. D. (1977). A technique of computer content analysis of transliterated Russian language textual materials: A research note. *American Political Science Review*, *71*(1), 245–251.

Casas, A., Denny, M. J., & Wilkerson, J. (2020). More effective than we thought: Accounting for legislative hitchhikers reveals a more inclusive and productive lawmaking process. *American Journal of Political Science*, *64*(1), 5–18.

Crawford, K., & Calo, R. (2016, October 13). There is a blind spot in AI research. *Nature*. https://www.nature.com/news/there-is-a-blind-spot-in-ai-research-1.20805

Dahl, R. A. (1961). The behavioral approach in political science: Epitaph for a monument to a successful protest. *American Political Science Review*, *55*(4), 763–772.

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, *26*(2), 168–189.

Diermeier, D., Godbout, J.-F., Yu, B., & Kaufmann, S. (2012). Language and ideology in Congress." *British Journal of Political Science*, *42*(1), 31–55.

D'Orazio, V., Landis, S. T., Palmer, G., & Schrodt, P. (2014). Separating the wheat from the chaff: Applications of automated document classification using support vector machines. *Political Analysis*, *22*(2), 224–242.

Easton, D. (1953). *The political system: An inquiry into the state of political science*. Alfred A. Knopf.

Easton, D. (1969). The new revolution in political science. *American Political Science Review*, *63*(4), 1051–1061.

Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E., & Stewart, B. M. (2018). How to make causal inferences using texts. arXiv:1802.02163.

Evans, M., McIntosh, W., Lin, J., & Cates, C. (2007). Recounting the courts? Applying automated content analysis to enhance empirical legal research. *Journal of Empirical Legal Studies*, *4*(4), 1007–1039.

Farr, J., & Seidelman, R. (1993). *Discipline and history: Political science in the United States*. University of Michigan Press.

Gentzkow, M., & Shapiro, J. M. (2010). What drives media slant? Evidence from US daily newspapers. *Econometrica*, *78*(1), 35–71.

Gentzkow, M., Shapiro, J. M., & Taddy, M. (2016). *Measuring polarization in high-dimensional data: Method and application to congressional speech*. Stanford Institute for Economic Policy Research. https://siepr.stanford.edu/sites/default/files/publications/16-028.pdf

Goet, N. D. (2019). Measuring polarization with text analysis: Evidence from the UK House of Commons, 1811–2015. *Political Analysis*, *27*(4), 518–539.

Grimmer, J.. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis*, *18*(1), 1–35.

Grimmer, J. (2013). Appropriators not position takers: The distorting effects of electoral incentives on congressional representation. *American Journal of Political Science*, *57*(3), 624–642.

Grimmer, J., Messing, S., & Westwood, S. J. (2012). How words and money cultivate a personal vote: The effect of legislator credit claiming on constituent credit allocation. *American Political Science Review*, *106*(4), 703–719.

Grimmer, J., Westwood, S. J., & Messing, S. (2014). *The impression of influence: Legislator communication, representation, and democratic accountability*. Princeton University Press.

Harris, J. A. (2015). What's in a name? A method for extracting information about ethnicity from names. *Political Analysis*, *23*(2), 212–224.

Herzog, A., & Benoit, K. (2015). The most unkindest cuts: Speaker selection and expressed government dissent during economic crisis." *The Journal of Politics*, *77*(4), 1157–1175.

Jamal, A. A., O Keohane, R., Romney, D., & Tingley, D. (2015). Anti-Americanism and anti-interventionism in Arabic Twitter discourses. *Perspectives on Politics*, *13*(1), 55–73.

Keith, K. A., Jensen, D., & O'Connor, B. (2020). Text and causal inference: A review of using text to remove confounding from causal estimates. arXiv:2005.00649.

King, G., & Lowe, W. (2003). An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization*, *57*(3), 617–642.

King, G., Pan, J., & Roberts, M. E. (2017). How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review*, *111*(3), 484–501.

Lang, S., Aull, L., & Marcellino, W. (2019). A taxonomy for writing analytics. *The Journal of Writing Analytics*, *3*, 13–37.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, *97*(2), 311–331.

Laver, M., & Garry, J. (2000). Estimating policy positions from political texts. *American Journal of Political Science 44*(3), 619–634.

Lowe, W. (2008). Understanding Wordscores. *Political Analysis*, *16*(4), 356–371.

Moxley, J., Elliot, N., Eubanks, D., Vezzu, M., Elliot, S., & Allen, W. (2017). Writing analytics: Conceptualization of a multidisciplinary field. *The Journal of Writing Analytics*, *1*, v–xvii.

Moxley, J., Elliot, N., Lang, S., Eubanks, D., Vezzu, M., Nastal, J., Tackitt, A., Phelps, J., & Osborn, M. J. (2019). Writing analytics: Broadening the community. *The Journal of Writing Analytics*, *3*, i–xi.

Owens, R. J., & Wedeking, J. P. (2011). Justices and legal clarity: Analyzing the complexity of US Supreme Court opinions. *Law & Society Review*, *45*(4), 1027–1061.

Pearl, J.. (2009). *Causality*. Cambridge University Press.

Peterson, A., & Spirling, A. (2018). Classification accuracy as a substantive quantity of interest: Measuring polarization in Westminster systems. *Political Analysis*, *26*(1), 120–128.

Popping, R. (2000). *Computer-assisted text analysis*. Sage.

Proksch, S.-O., Wratil, C., & Wäckerle, J. (2019). Testing the validity of automatic speech recognition for political text analysis. *Political Analysis*, *27*(3), 339–359.

Prothro, J. W. (1956). Verbal shifts in the American presidency: A content analysis. *American Political Science Review*, *50*(3), 726–739.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, *54*(1), 209–228.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*(4), 1064–1082.

Rodman, E. (2020). A timely intervention: Tracking the changing meanings of political concepts with word vectors. *Political Analysis*, *28*(1), 87–111.

Rubin, D. B. (2006). *Matched sampling for causal effects*. Cambridge University Press.

Schattschneider, E. E. (1975). *The semisovereign people: A realist's view of democracy in America*. Wadsworth Publishing Company.

Schubert, G. A. (1958). The study of judicial decision-making as an aspect of political behavior. *American Political Science Review*, *52*(4), 1007–1025.

Seidelman, R. (2015). *Disenchanted realists: Political science and the American crisis*. SUNY Press.

Silge, J., & Robinson, D. (2016). Tidytext: Text mining and analysis using tidy data principles in R. *Journal of Open Source Software*, *1*(3), 37.

Soroka, S., Daku, M., Hiaeshutter-Rice, D., Guggenheim, L., & Pasek, J. (2018). Negativity and positivity biases in economic news coverage: Traditional versus social media." *Communication Research*, *45*(7), 1078–1098.

Spirling, A. (2012). U.S. treaty making with American Indians: Institutional change and relative power, 1784–1911. *American Journal of Political Science*, *56*(1), 84–97.

Spirling, A. (2016). Democratization and linguistic complexity: The effect of franchise extension on parliamentary discourse, 1832–1915. *The Journal of Politics*, *78*(1), 120–136.

Tetlock, P. E. (1981). Personality and isolationism: Content analysis of senatorial speeches. *Journal of Personality and Social Psychology*, *41*(4), 737.

White, H. B. (1956). Commentary on Prothro's content analysis. *American Political Science Review*, *50*(3), 740–750.

Wilkerson, J., Smith, D., & Stramp, N. (2015). Tracing the flow of policy ideas in legislatures: A text reuse approach. *American Journal of Political Science*, *59*(4), 943–956.

Young, L., & Soroka, S. (2012). *Lexicoder Sentiment Dictionary*. http://www.lexicoder.com