Text as Data in Interest Group Research

Aizenberg, E.

DOI
10.1007/978-3-030-13895-0_31-1

Publication date
2020

Document Version
Author accepted manuscript

Published in
The Palgrave Encyclopedia of Interest Groups, Lobbying and Public Affairs

Link to publication

Citation for published version (APA):
Aizenberg, E. (2020). Text as Data in Interest Group Research. In P. Harris, A. Bitonti, C. S. Fleisher, & A. Skorkjær Binderkrantz (Eds.), *The Palgrave Encyclopedia of Interest Groups, Lobbying and Public Affairs* (Living ed.). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-13895-0_31-1

Text as Data in Interest Group Research

Author
Ellis Aizenberg (e.aizenberg@uva.nl), Department of Political Science, University of Amsterdam

Keywords

Text analysis, Computational methods, Interest groups

Definition

Text analysis is highly useful for scholars working on lobbying and interest groups. It is a method through which organized interests, their positions on issues and the frames that they employ can be identified within bodies of text. This technique allows scholars to measure access to different types of venues and the policy positions that groups express. It can reveal patterns of populations and communities that can be used as (in)dependent variables. In turn, these can be linked to characteristics of organized interests, the context in which they operate and policy outcomes.

Analysis of content in texts can be best described as 'the systematic, objective, quantitative analysis of message characteristics'(Neuendorf, 2002: 1). There are different types of text analysis and these types can usually best be described by the degree of human involvement in the coding process. The types range from human coding to supervised coding to unsupervised coding which are associated with different trade-offs when it comes to validity and reliability.

This entry explains why text analysis is a useful method for interest group scholars, sheds light on the different types of text analysis, their benefits and limitations, notes the main empirical applications in the field of interest group research and discusses promising applications for future research.

Introduction

Studying the creature that is referred to as the interest group can be a challenging task. That is, interest groups operate in high numbers and often in relative secrecy, compared to political parties or political leaders. Interest group scholars therefore oftentimes have to be creative when they seek to study this phenomenon empirically.

Text analysis has been employed by several scholars – benefiting from the amount of available text that is vast and growing. Examples are the increasing online availability of transparency registers, press releases, minutes of parliamentary meetings, policy documents and newspaper coverage. These documents can be used for text analysis and help scholars answer questions about patterns of interest group populations, access and influence (see chapters on interest group populations, interest group access and influence).

While surveys and interviews have enlightened us with important findings concerning the activity and impact of organized interests, text analysis can help to answer questions of a slightly different nature. That is, both surveys and interviews rely on data that is self-perceived while documents allow for measurement of observational data. What is more, it grants us access to view into the empirical world of the past.

This entry proceeds with a discussion of the different types of text analysis and the associated trade-offs. Subsequently, it elaborates on the empirical studies within the field that have employed text analysis and illustrates some interesting applications for future research.

Text analysis and trade-offs

Text analysis can best be characterized by the degree of human involvement and varies from classic content analysis to unsupervised machine learning. Full manual coding can be carried out by employment of a pre-defined coding scheme. It is associated with high levels of validity and is technically easy to employ. It can however, become procedurally complex as it often requires training of coders. Another advantage is that humans are able to understand phenomena from texts that are highly complex. The main disadvantage related to hand-coding is that there are reliability issues related to this approach. Problems with reliability can especially occur when the coding task at hand is ambiguous and when there are multiple human coders involved as different coders could potentially understand the coding task in different ways. Reliability of manual coding can be evaluated on the basis of intercoderreliability tests and measures such as the Krippendorff's alpha (see Krippendorff, 2004 for an elaborate discussion on different coefficients that exist to assess reliability). Hand-coding is also time-consuming and consequently requires a sufficient budget.

A second category of text analysis – computer assisted analysis – is characterized by less involvement of humans. Examples of methods that require the involvement of both humans and computers are querying and dictionary coding. Key words are central to both methods. Scholars can employ querying when interested in knowing how often a certain word appears in a political text and dictionary coding when interested in broader phenomena. Dictionary coding relates certain words to categories that are pre-defined (see Laver & Garry, 2000 for an empirical example that estimates policy positions in political texts). Such methods employ the relative frequency of certain key words to measure categories (Grimmer & Stewart, 2013). Another range of methods that belongs to this category are supervised machine learning techniques. Here, humans code a set of documents manually. The algorithm learns from the hand-coded dataset and applies the coding to the remainder of the documents. These methods can be validated with statistics of model performance and ensure scholars to create and implement clear concepts in their coding endeavors (see Grimmer & Stewart, 2013 for an elaborate discussion and Fraussen et al., 2018 for an application in the field). An example of such a technique is known entity recognition (NER) which is a method that identifies and extracts entities such as events, people and places from a text corpus. NER applications based on dictionaries or unsupervised machine learning also exist (see Nadeau & Sekine, 2007 for an elaborate overview and discussion).

A third category can be placed on the other end of the continuum when it comes to human involvement. That is, unsupervised machine learning. This category refers to a "class of methods that learn underlying features of text without explicitly imposing categories of interest" (Grimmer & Stewart, 2013:281). Thus, for such methods it is not necessary to have a pre-defined set of categories one is interested in. Applying unsupervised learning techniques can be useful when one seeks to study a text corpus but is unsure which categories are of interest. An example is topic modelling such as latent dirichlet allocation (see both Blei et al., 2003 and Grimmer & Stuart, 2013 for more information and elaborate discussions on the method). Both R and Python are recommended for the management and analysis of text analysis. Suitable packages are *corpustools* and *quanteda* (see Welbers et al., 2017 and Benoit et al., 2018 for further reading).

A major advantage associated with the usage of computers is that they are able to count quickly and can therewith work with large datasets in a relatively short period of time. Computers however, are able to count, yet cannot read texts as humans can. It is therefore of utmost importance to carry out extensive validity tests (see both Grimmer & Stewart, 2013 and Klüver, 2015: 460 for an elaborate discussion). It is furthermore also important to note that any form of text analysis requires substantial effort to collect, preprocess and store the text documents of interests. Data collection can be done through manual downloading, bulk downloading or through data scraping. The latter technique is an automated technique which reads, extracts and stores data in a desired structure and place and can be carried out through R or Python. When the data of interest is collected, there is often quite some work to be done to preprocess the text in order to reduce dimensionality. Examples are the removal of stop words and punctuation.

Applications of text analysis in interest group research

Text analysis has been applied by several scholars in the field of interest group research. Ever since Schattschneider (1960) coined the term of bias in the pressure system, political scientists have been struggling with this concept both theoretically and empirically. That is, there is no yardstick against which to assess bias (see Lowery et al., 2015). Reflecting the call to study bias in the pressure system empirically, a strand of research has focused on assessing the appearance of organized interests in different political arenas. Binderkrantz (2012) for example, has assessed the diversity of interest group media attention in the Danish news media between 1984 and 2003. She shows that diversity has increased over time as both public and sectional groups gained more relative attention over time. The coding approach of this study is fully manual.

Others apply text analysis as well to study patterns of access. For instance, Aizenberg and Hanegraaff (2020) employ data scraping and a query based method to identify the organized interests that gain access to parliamentary meetings in the Netherlands between 1970 and 2017. They find that corporate access to parliamentary hearings has increased and by employing it as a dependent variable, they show that it is triggered by economic downturns and new political opportunities. Fraussen and colleagues (2018) employ a query based method as well to identify organizations in parliamentary debates in Australia. In order to assess ‘prominence’ – operationalized as groups that are used as a resource by legislators – the authors use a supervised machine learning technique through which they show that the pattern of prominence is highly skewed to a small number of groups.

Text analysis is also applied by scholars interested in measuring the degree to which interest group policy positions and policy outcomes overlap. Klüver (2009) presents a case study in which she codes both texts of interest groups and the European Commission in the consultation process. She compares the results of hand-coding to the results of two automatic approaches, a supervised machine learning technique and an unsupervised machine learning approach. The results strongly correlate and therewith demonstrate that the techniques can be useful tools to measure overlap between policy positions of interest groups and outcome. Similarly, Bunea and Ibenskas (2015) compare estimates of qualitative hand-coding and those of an unsupervised scaling approach in the same context, but, in contrast to Klüver, find little correlation between the two. The authors criticize ‘Bag of Words’ approaches in quantitative text analysis as they neglect the broader context in which these words occur as well as the validity of an unsupervised machine learning approach, leading them to argue that employing unsupervised learning algorithms raises challenges when it comes to analyzing lobbying in the European Union.

In a response, Klüver (2015) addresses the issues raised by Bunea and Ibenskas (2015). She phrases that it is indeed a ‘simplifying assumption’ as language is contextual but that this does not mean that such analysis cannot provide valid results (Klüver, 2015:457). She emphasizes that quantitative text analysis techniques can be promising as long as scholars pay careful attention to the assumptions of the models and rely on both internal and external validation techniques such as hand-coding and cross-validation of the output with estimates of other quantitative text approaches as well as external substantive data (Klüver, 2015:460).

Conclusion

Text analysis is a method through which one can systematically analyze characteristics of content.

Political scientists studying organized interests and lobbying, have used text documents as data sources to study interest group communities in different political arenas and linked policy positions of interest groups to policy outcomes. In doing so, they have benefited from the increasing availability of text corpora and have seized the opportunity to rely on observational data and the ability to shed light on the empirical world of the past. Interest group scholars which have employed text analysis for their endeavors have contributed to our knowledge in many ways, most notably on the shapes of interest group communities and preference attainment. Others have carefully shown that scholars need to be cautious when using these techniques as they can produce invalid findings. It is therefore of utmost importance that the researcher always carries out extensive internal and external validity checks on the produced findings.

Although caution is thus highly desired when relying on text analysis, it holds promising paths for future academic research. To this date however, text methods are not applied to their full potential in the field. With the right skill set, every political scientist working on lobbying and interest groups, can exploit the vast amount of text that is available. A first avenue for future research that is worth highlighting builds on the work on prominence that has been carried out by Fraussen and colleagues (2018). In their endeavor, they sought to not only identify stakeholders that participate in politics but also account for the context in which they do so in an automated manner. A different way to involve context, would be to apply clause analysis. This method automatically grasps the context in which political actors operate by extracting actions and statements of actors in an automatic manner from a text document (see Van Atteveldt et al., 2017).

A second interesting path for future studies would be to focus on the dynamics displayed on social media. An interesting example is set by Theocharis and colleagues (2016) who assessed Twitter's democratic promises by studying the interaction between party candidates and citizens through employing automated text analysis. Studying the interaction between organized interests, their members and political elites are all promising avenues for future research.

Cross references

Interest group Access/Interest group populations/Influence

References

Aizenberg, E., & Hanegraaff, M. (2020). Is politics under increasing corporate sway? A longitudinal study on the drivers of corporate access. *West European Politics*, 43(1), 181–202.

Binderkrantz, A. S. (2012). Interest groups in the media: Bias and diversity over time: interest groups in the media. *European Journal of Political Research*, 51(1), 117–139.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.

Bunea, A., & Ibenskas, R. (2015). Quantitative text analysis and the study of EU lobbying and interest groups. *European Union Politics*, 16(3), 429-455.

Fraussen, B., Graham, T., & Halpin, D. R. (2018). Assessing the prominence of interest groups in parliament: a supervised machine learning approach. *The Journal of Legislative Studies*, 24(4), 450-474.

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(03), 267–297.

Klüver, H. (2009). Measuring Interest Group Influence Using Quantitative Text Analysis. *European Union Politics*, 10(4), 535–549.

Klüver, H. (2015). The promises of quantitative text analysis in interest group research: A reply to Bunea and Ibenskas. *European Union Politics*, 16(3), 456–466.

Krippendorff, K. (2004). Reliability in content analysis. *Human communication research*, 30(3), 411-433.

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

Lowery, D., Baumgartner, F. R., Berkhout, J., Berry, J. M., Halpin, D., Hojnacki, M., … Schlozman, K. L. (2015). IMAGES OF AN UNBIASED INTEREST SYSTEM. *Journal of European Public Policy*, 22(8), 1212–1231.

Nadeau, D., Sekine, S., 2007. A Survey of Named Entity Recognition and Classification. *Lingvisticae Investigationes* 30, 3-26.

Neuendorf, K. A. (2002). *The content analysis guidebook*. Sage.

Schattschneider, E.E. (1960). *The semisovereign people: A realist's view of democracy in America*. New York: Holt, Reinhart & Winston.

Further Reading

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. (2018) "quanteda: An R package for the quantitative analysis of textual data". *Journal of Open Source Software*. 3(30), 774.

Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S. A., & Parnet, O. (2016). A bad workman blames his tweets: the consequences of citizens' uncivil Twitter use when interacting with party candidates. *Journal of communication*, 66(6), 1007-1031.

Van Atteveldt, W., Sheafer, T., Shenhav, S. R., & Fogel-Dror, Y. (2017). Clause analysis: Using syntactic information to automatically extract source, subject, and predicate from texts with an application to the 2008–2009 Gaza War. *Political Analysis*, 25(2), 207-222.

Welbers, K., Atteveldt, W. V., & Benoit, K. (2017). *Text Analysis in R. Communication Methods and Measures*, 11(4), 245–265. https://doi.org/10.1080/19312458.2017.1387238