Text-as-Data Methods for Comparative Policy Analysis*

Fabrizio Gilardi† Bruno Wüest‡

November 5, 2018
7,611 words

1 Introduction

Text-as-data methods are a broad set of techniques and approaches relying on the automated or semi-automated analysis of text. They have become increasingly prevalent in the social sciences, and are part of a broader trend in which, taken together, the internet and computational social science tools have changed the kinds of questions that social scientists can ask and answer successfully (Golder and Macy, 2014; Lazer and Radford, 2017). Text analysis holds a prominent place in these developments. Texts have always been a primary data source for social scientists. As Monroe and Schrodt (2008, 351) write, "text is arguably the most pervasive—and certainly the most persistent—artifact of political behavior." In the internet age, texts have become particularly plentiful, and accessible with relative ease. The large amount of text available to researchers, combined with new computational tools, have promoted the development of text-as-data approaches in which texts are analyzed statistically with different degrees of automatization. The promise of the approach is that it can both apply existing theories to new data and uncover new phenomena that previously remained hidden (Evans and Aceves, 2016). As González-Bailón (2017, xviii) writes: "when the right connections are made, much of the data-driven research that is being conducted today speaks directly to long-standing (and unresolved) theoretical discussions."

Text-as-data approaches are becoming mainstream in political science. Typical applications revolve around research question where at least one element is based on political communication theories such as agenda setting, issue definition, or framing (for reviews, see Grimmer and Stewart, 2013; Lucas et al., 2015; Wilkerson and Casas, 2017). From a practical perspective, these approaches allow researchers to conduct more efficiently research they have been doing

*We thank Christian Breunig, Guillaume Fontaine, and B. Guy Peters for helpful comments.
†Department of Political Science, University of Zurich, https://www.fabriziogilardi.org/
‡Department of Political Science, University of Zurich, https://www.bruno-wueest.ch/

manually for decades, such as classifying texts into categories such policy areas. But, thanks to powerful inductive analysis procedures, text-as-data approaches also make it possible to discover new phenomena, concepts, and correlations from latent dimensions of texts.

The focus of this chapter is on comparative policy analysis more specifically, a field in which text-as-data methods have not yet been applied widely despite the potential demonstrates by applications in other subfields. Textual materials covering various administrative, legislative, and political aspects of policy processes have always been a central data source for policy analysis. For example, texts have been the central source for the successful and influential Comparative Agendas Project (John, 2006; Baumgartner et al., 2006, 2011; Dowding et al., 2016), which originally relied on manual coding to classify legislation and other relevant texts into 21 major topics and 220 subtopics,[1] but is increasingly using automated approaches to carry out this task.

The specificity of text-as-data-methods for comparative policy analysis relies in the application of existing tools and approaches to both classic and new theories and phenomena relevant for public policy and policy analysis. However, it is possible, and certainly desirable, that scholars of policy analysis will build on existing methods to adapt them to their specific research needs.

The goal of the chapter is to offer an overview of existing applications and, especially, of the options and workflow of text-as-data approaches for comparative policy analysis. The learning curve can at first appear quite steep for these methods, and we hope to motivate policy analysts to take on the challenge by showing the potential payoffs as well as offering an overview of the various practical steps and available options. To this end, the chapter first reviews applications of text-as-data methods in comparative policy analysis. Then, it focuses on the practical aspects of these methods, and specifically on the workflow involved in their application, such obtaining and storing the data, pre-processing, and analyzing them with a range of automated and semi-automated techniques. Next, it presents three specific kinds of applications: concept identification, classification, and discovery. The chapter concludes by highlighting the potential of text-as-data methods for comparative policy analysis despite their relatively sparse use so far.

2 Text-as-Data Applications in Comparative Policy Analysis: An Overview

The use of text-as-data techniques in political science has increased steadily in the past years and has become highly diversified. The origins of text-as-data methods can be tracked to a various approaches such as classical content analysis (Krippendorf, 2004) and the computer science literature on natural language processing (Jurafsky and Martin, 2009). In political science, event detection systems for conflict studies are among the earliest applications (Gerner et al., 1994).

[1] `https://www.comparativeagendas.net/pages/master-codebook`, accessed March 7, 2018.

Originally, these approaches used keyword matching to encode events such as bomb attacks. Later, these approaches were enhanced with more advanced tools such as syntax parsing, lexical databases, and named entity recognition tools. Other early applications in political science relied on scaling approaches, which involves a broad array of methods aiming to map texts on one or more underlying dimensions. Starting with the Wordscore method (Laver et al., 2003; Lowe, 2008) and the widely-used Wordfish method (Slapin and Proksch, 2008), these approaches have been used primarily to extract ideological or policy-specific ideal points from texts such as legislative speeches and party manifestos.

In the last decade, a rapidly growing literature has aimed to develop tools for classifying political texts (e.g., Hopkins and King, 2010). Most commonly, politically meaningful texts are classified into issue or topic categories. Such classifications can be conducted deductively using machine learning, or it can be conducted inductively with latent variable models. Examples include studies that analyze policy-specific debates in parliaments (Quinn et al., 2010), censorship by authoritarian regimes (King et al., 2013), electoral representation (Grimmer, 2013), and individual opinion and decision-making (Wüest, 2018). The rise of social media platforms such as Facebook and Twitter for political communication has further opened the way for social network and big data analysis into political science (for an overview, see Jungherr and Theocharis, 2017).

In comparative policy analysis, applications of text-as-data methods are relatively rare, despite the potential demonstrated by their increasing popularity in other related fields. To be sure, qualitative text analysis is prevalent in the literature, sometimes in combination with quantitative methods. For example, the discourse networks approach relies on text analysis to measure discourse coalitions quantitatively through network analysis (Leifeld and Haunss, 2012; Leifeld, 2013; Fisher et al., 2013b,a). Although text-as-data applications are not yet mainstream in comparative policy analysis, a few notable exceptions have addressed questions that have been at the core of policy analysis for many years.

Policy agendas. The Comparative Agendas Project is a network of researchers developing a measurement system to classify a broad range of political activities into topics which can be compared over time and across political systems (Baumgartner and Jones, 2018). The project builds on Baumgartner and Jones' pioneering work on policy agendas (Baumgartner and Jones, 1993; Jones and Baumgartner, 2005). Initially, and for many years, the project employed on manual coding to classify a wide range of legislative, judiciary and journalistic texts into policy categories, using a very detailed coding scheme. More recently, the project has started to rely on machine-learning procedures in which manually prepared training data sets are fed to classification algorithms (Collingwood and Wilkerson, 2012). The move to automatized procedures was needed particularly in countries or areas that were not part of the original project or for which the costs of manual coding were prohibitive, such as Hungary (Sebók and Berki,

2017) and Danish city councils (Loftis and Mortensen, 2018). These examples illustrate one of the main advantages of text-as-data methods, namely, the possibility to extend existing projects to new areas at a relatively low cost.

Problem definition. A textbook argument in policy analysis is that the decision-making process consists of several stages, and that the stage in which problems are defined as politically relevant is a crucial one. As Elder and Cobb (1984, 115) noted, because "policy problems are not a priori givens but rather are matters of definition [...] what is at issue in the agenda-building process is not just which problems will be considered but how those problems will be defined." The idea that problem definition affects the kinds of policies that are adopted, as well as those that are not, is now considered "nearly axiomatic" within the policymaking literature (Boushey, 2016, 200). Texts are a natural source to study problem definition, and some studies have started to rely on text-as-data approaches to do so. Nowlin (2016) shows how topic models can be used to study how issues are defined and applies the approach to Congressional hearings regarding used nuclear fuel. His analysis identifies seven dimensions (programmatic, safety/regulation, Yucca mountain, site selection, science/technical, storage, and transportation) and shows that their salience co-varies with important policy developments. Gilardi et al. (2018) apply topic models to newspaper articles on smoking bans in US states and find significant differences in how the issue was defined, both across states and over time. They employ structural topic models to extract the issue definitions and, at the same time, to estimate the correlation of the issues with covariates such as the sentiment with respect to smoking bans. The Policy Frames Project, finally, uses machine learning to track media tone and framing in a variety of areas (Card et al., 2015, 2016), with the aim of providing a comprehensive scheme for the identification of policy frames. This example illustrates another advantage of the text-as-data approach, namely, the possibility to build on other researchers' work at sharply decreasing marginal costs.

Policy diffusion. Policy diffusion refers to the phenomenon whereby policies in one unit (city, state, country, etc.) are influenced by policies in other units (Simmons et al., 2006; Braun and Gilardi, 2006; Graham et al., 2013). It is a classic question going back at least to Walker (1969), which in the policy analysis literature is also studied under the label "policy transfer" Dolowitz and Marsh (1996); Dolowitz (2000). Traditionally, the focus has been on policy adoptions, but scholars have increasingly been interested in how diffusion affects other aspects of the policy-making process. For example, Gilardi et al. (2018) study how issue definitions diffuse across US states, finding that practical aspects of smoking restrictions are more subject to diffusion than normative rationales. Wilkerson et al. (2015) use a text reuse approach to trace how ideas spread from one piece of legislation to the other, as well as from initial drafts to final bills, which helps uncover the influence of specific lawmakers. Linder et al. (2018) show how the approach can be used to measure policy similarity more in general, especially in a diffusion context. These works illustrate how text-as-data methods can be used to study classic questions

from new angles that were previously impractical due to technical constraints.

Lobbying. Klüver (2013) applies Wordfish (Slapin and Proksch, 2008), a scaling method to reduce the dimensionality of texts, to measure the policy preferences of interest groups based on their submissions in online consultation of the European Commission. The analysis underscores the collective nature of lobbying, where success depends on the interaction between information supply, citizen support, and economic power of lobbying camps. Again using consultations, Klüver et al. (2015); Klüver and Mahoney (2015) use cluster analysis to identify the frames used by interest groups as well as their determinants and effectiveness. Klüver et al. (2015) shows how interest groups tailor their frames based on their targets, while Klüver and Mahoney (2015) show how the European Commission has adopted the frames put forward by various lobbies. This streams of research shows how text-as-data methods have been integrated in a well-established literature.

Policy feedbacks. The feedback effects of policies, and specifically how policies affect political dynamics, is a classic question Pierson (1993) that is crucial to path dependence theories Pierson (2000). Flores (2017) studies this question using a dictionary approach to measure the sentiment of tweets to identify how public opinion reacts to anti-immigrant legislation. The study finds that the policy affected public discourse not by changing attitudes, but by mobilizing people already critical of immigrants. This is again an example of how text-as-data methods can shed new light on long-standing discussions in policy analysis.

Most of these approaches have in common that they rely on statistical algorithms that estimate the quantities of interests from a bag-of-words representation—basically word frequency distributions—of the texts. This also implies an application of basic concepts of statistical learning such as training and cross-validation. In addition, most of the text-as-data applications discussed above use large or even very large text collections that cannot be analyzed manually with reasonable effort. Another characteristic of these studies is that they rely on existing approaches, which they apply to a specific research question. I other words, text-as-data applications in comparative policy analysis take advantage of methods developed in other subfields. Policy analysis adjust the methods to their specific needs, but they have not not created new text-as-data approaches as such.

3 Text-as-Data in Practice

Text-as-data approaches involve many moving parts and can be highly complex. On the side of the prediction, applications of automated text analysis rely primarily on statistical or heuristic algorithms that retrieve information from bag-of-words representations of the original texts. That is, many kinds of information conveyed in text documents, from morphological information such as word order and word ambiguities to semantic information such as irony, metaphors

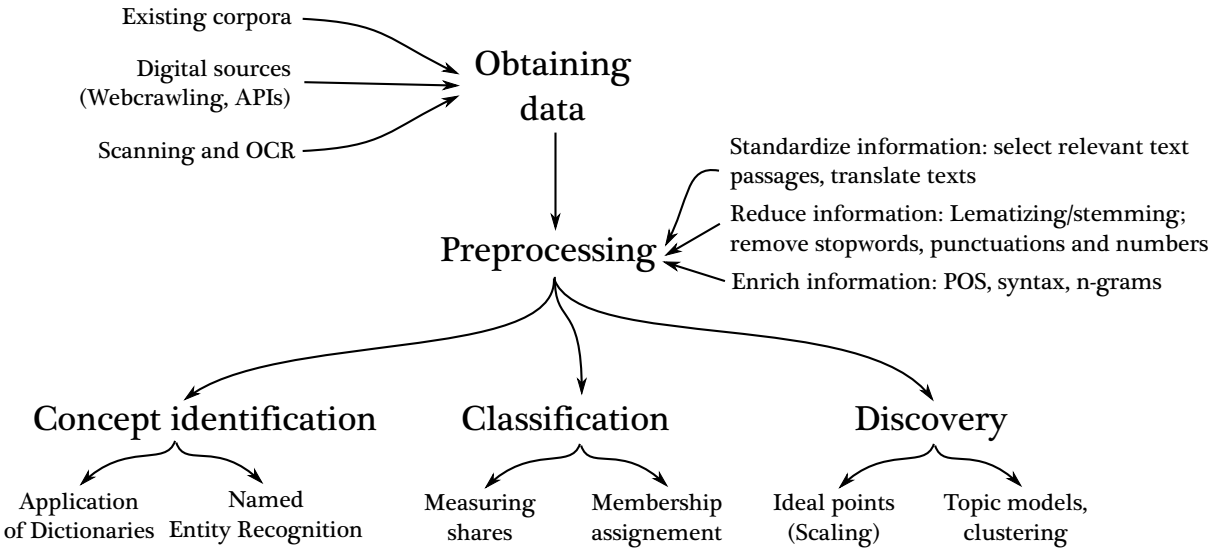

Figure 1: *Overview of text-as-data applications.*

or double negatives, are often not discarded in the pre-processing steps prior to the analysis. Although such simplifications can appear unreasonable, they have been shown to be surprisingly effective for many kinds of applications. However, they also imply that all results of automated models of language at best are useful approximations of the quantities of interest (Grimmer and Stewart, 2013). Moreover, they imply that automated models of language necessarily are highly domain-specific. Accordingly, there are no globally best methods for retrieving certain information from texts. A careful selection and validation of both the preparatory steps and the methods of analysis are therefore essential for the success of a research project. Whenever possible, the validity of text-as-data applications should always be quantified and reported using commonly accepted measures such as *recall*, *precision* and Kohen's κ.

Figure 3 shows the main steps of a text-as-data application, which will be described more in detail in the following sections. A first aspect that is often overlooked are the procedures to obtain the data. Then, researchers need to decide how the documents should be pre-processed. Finally, different research goals imply different methodological strategies, so researchers need to make the theoretically and practically appropriate choices in terms of the methods applied.

The following discussion will first present the preparatory steps in Section 3.1, before the various methods are outlined in Section 3.2.

3.1 Preparing a Text Corpus

3.1.1 Obtaining the Data

The unheralded first step in text-as-data projects, the construction of a text corpus, usually requires a lot of effort. This is because many documents of interest for researchers in public policy

are not easily accessible. For instance, it may come as a surprise that about a third of the state legislatures in the United States do not regularly publish their floor debates on the Internet. In some cases, it is simply a question of accessibility, and can be resolved with contacting the database provider. In most cases, however, there do not even exist digital versions of the documents of interest. More often than not, historical archives have not been digitized, and it is often these archives that are of particular importance to do diachronic policy analysis. In this scenario, the only option is to climb into the archives, to scan the documents and run an Optical Character Recognition (OCR) software on the scans. Depending on the quality of the documents – e.g. the fonts' sharpness and the contrast between fonts and sheets –, this step can already require a lot of effort.

In addition, the terms and conditions of many database providers are often all but conducive to large-scale text mining projects. First, the usual web interfaces mostly do not allow bulk downloads of texts, either because it is explicitly prohibited or because web scraping is too slow to retrieve a large number of documents within a reasonable time frame. In other words, if the search and download in a web interface takes several seconds and only a few texts can be downloaded at once, it easily takes months until a large corpus is compiled. Scraping from Possible solutions are programmable interfaces (API) to the providers' database or that a special agreement for a one-time transfer of large data can be negotiated. Second, text mining on the retrieved documents often is prohibited. Most newspapers, for example, explicitly prohibit automated information extraction from their articles in their terms and conditions. The same holds for the fact that most original data from commercial database providers cannot not be published. Obviously, this runs contrary to both the scientific principle of reproducibility and the open data policy of many publishers. In practice, this is an extensive gray area, with one position arguing that almost everything can and should be published anyway – e.g. that documents with one word removed from the text do not count as original data anymore – and another position only using where an explicit exception has been granted by the database provider (Tennant et al., 2016). To mitigate any uncertainties, approaching the database providers and transparently negotiating the terms of analysis and publication seems the most promising way. Here, university bodies such as the central library can provide essential support.

As for the storing of the data, we would recommend a different infrastructure depending on the capabilities and resources available to the researchers. Optimally, a large text corpus should be stored in non-relational database such as *ElasticNet* or *MongoDB*, which allow for an efficient storage as well as fast document searches (Jatana et al., 2012). This set-up, however, needs particular IT-skills and a server infrastructure that may not be available in every research team in comparative policy analysis. An alternative, low-threshold solution is the storage of a corpus in single text files that are a systematically stored in a folder tree (e.g. organized by source and date of publication). This means longer times to load the corpus, but it may be easier to keep

an overview of the data for researchers not used to work with large text data.

3.1.2 Preprocessing

Given the inherently unstructured nature of text data, it needs careful preparation before it can be analyzed. The preprocessing of the text data involves three steps and can be carried out using several tools such as `spacyR`, `TM`, `quanteda` or `udpipe` in `R`, or `NLTK`, `spacy` and `polyglot` in `python`.

First, when preparing a corpus, researchers need to invest in the standardization of the texts. Several technical details such as the standardization of character encodings and the extraction of meta-data (publication dates, authors etc.) need to be clarified. Especially if the documents stem from different sources, the encoding may vary depending on the operating systems and software programs used to process these documents. We recommend to standardize texts into one of the most common encodings suitable also for special characters such as German umlauts, such as *utf8* or *latin1*. Meta-data, on the other hand, include information, such as the source or the author of the documents, that can be very important for the analysis. Standardization also involves the definition of the units of observation, that is, the relevant text passages. For example, if parliamentary speeches are analyzed and members of parliament (MP) are the main subjects of study, the full speeches can be defined as the unit of observation. If the same MPs are to be analyzed in newspaper articles, in contrast, it can be helpful to restrict the analysis to the paragraphs mentioning the MPs, leaving paragraphs discussing other topics aside.

Second, a crucial aspect to consider during preprocessing is that most automated text analysis applications are language-specific. If documents in more than one language are to be included in the same analysis, they can either be translated into one language and then analyzed by one single model, or they should be analyzed with separate models. The former has the advantage that one result is estimated that holds for the whole corpus. Some semantic nuances of the texts such as emotions, however, can be lost during the translation. The latter, in contrast, suffers from the problem that the results produced by the different models may not be straightforwardly comparable. de Vries et al. (2018) have shown that, for many text-as-data applications, machine translation performs almost as well as expert translation. Therefore, researchers should seriously consider this option when working with multi-lingual corpora. Moreover, the quality of automatic translation is increasing steadily, with services such `DeepL`[2] being currently the state of the art.

Third, the information in the texts needs to be reduced by removing "stop words" (that is, common words such as "and" or "the"), punctuation, and numbers. The rationale to reduce this information is that not all text elements contain important information for the word distributions used in the estimations. Stemming or lemmatizing is another procedure to reduce

[2] `https://www.deepl.com/translator`.

complexity. Stemming cuts word endings, while lemmatizatizing transforms each word into its basic form. For most languages other than English, lemmatization should be preferred since there are many irregular conjugations and declinations. Researchers can also opt to enrich the text data by identifying the part-of-speech (POS) of words (that is, whether a given word is a verb, a noun, etc.), building n-grams (combinations of words), and extracting information on the syntactical dependencies of the words in the texts. Such methods can prove very useful to analyze short texts for which simple word distributions entail not enough variance. Since they add more layers of basic information such as the word order, they can considerably improve the estimations in some scenarios. All these decisions depend on the specific goals of the analysis as well as on the nature of the texts. Punctuation, for example, may only add noise for most estimations, but they have been found to be useful in classification of emotions. Because of this uncertainty, it is generally recommended to consider several alternatives and test empirically whether they improve the estimations.

The preprocessing of the text data has a decisive influence on the results (Denny and Spirling, 2017). Therefore, is is recommended to either extensively test the influence of every preprocessing step or, even better, to include the preprocessing parameters in the analysis procedures.

3.2 Methods

Text-as-data applications are one of the most dynamic areas of political science methodology. In the last years, contributions from this area have steadily increased. As the range of applications grows, it becomes difficult to keep track of all developments. The subsequent discussion tries to provide a broad overview over the area of text-as-data methods and applications. More precisely, we suggest that the many different applications can be grouped according to three different research goals: extraction of specific information (concept identification), theory-driven allocation (classification) and inductive exploration of the underlying dimensionality (discovery).

3.2.1 Concept Identification

The goal of concept identification is to find and extract the specific text passage that refers to a concept of interest. Concepts can be highly complex, such as the degree of conflict in a political speech, but also more straightforward, such as the names of governors of US states in press releases. In broad terms, concept identification methods can be separated into dictionary-based approaches and named entity recognition approaches. Applications of dictionaries, also referred to as "ontologies", "lists," or "gazetteers" in some literatures, involve matching keywords in the texts of interests. They are sometimes criticized for their simplicity. However, if the operationalization produces a comprehensive set of keywords that can be mapped unambiguously

to a concept, such approaches are highly reliable and efficient. A good example are names of politicians or political parties, which are quickly compiled and mostly refer unequivocally to the actors under concerns (see Wüest et al., 2016; Müller, 2015; Gilardi and Wueest, 2017). Sentiment analysis traditionally was also conducted using dictionaries of word polarities (Young and Soroka, 2012). Although these approaches are being increasingly replaced by supervised classifications using machine learning, they keep being actively developed, for instance for applications in multilingual settings (Proksch et al., 2018). Technically, the matching of dictionaries can be implemented using regular expressions (`regex`), which can be implemented using librarries that are part of the base distribution of both `R` and `python`.

A more complex set of methods for concept identification is usually termed Named Entity Recognition (NER). NER approaches are based on machine learning, which means that specific concepts are recognized by a model using linguistic rules and bag-of-words information from the word contexts of these concepts. There are NER tools that are trained on such large corpora such as Wikipedia sites (e.g. the Stanford NER or `polyglott` library in `python`) that they can be applied off the shelf. Hence, no dictionary has to be build when using NER tools, but they are usually only able to detect a restricted set of concepts such as persons, locations, dates or organizations. However, these are usually the concepts researchers in comparative policy analysis are interested in. The detection of locations, for example, can be used to assign documents to geographical units that are the subject of policy diffusion studies (Gilardi et al., 2018; Ciocan and Wueest, 2017).

3.2.2 Classification

Supervised classification tasks can be defined as a separate set of text-as-data methods. Text classification can either be used to assign class memberships, e.g. to which policies the documents of interest can be categorized, or to estimate class shares in documents, e.g. the relative importance of different policy debates in the same documents (Jurafsky and Martin, 2009; King et al., 2013).

The approach is similar for most application. First, a training set needs to be build, which usually involves the manually coding of a sample of the data to be classified. Increasingly, researchers in political science also use crowd-sourcing to build these training sets (Benoit et al., 2016). Then, a generative model is created and optimized. It uses the hand-coded inputs to calculates the probability that each document belongs to a certain category. Popular algorithms implementing such models are the multinomial naïve Bayes (Conway and White, 2012), support vector machines (Meyer, 2012), regularized paths for generalized linear models (Friedman et al., 2010) and maximum entropy (Jurka et al., 2013). Such models can be further optimized using bootstrapped training and cross-validation, evaluating the best trade-off between false positives and false negatives (which is also known as as optimization of the receiver operator characteris-

tic, or ROC), and building ensembles, that is, classifiers that include several algorithms or models and perform classification by comparing their predictions.

In many text-as-data applications classifications a are necessary first steps in order to compile the corpus of interest, since text data collection may contain a large number of false positives, that is, texts that do not actually belong in the corpus (e.g. Wüest et al., 2016; Gilardi et al., 2018; Ciocan and Wueest, 2017).

Software tools that allow several kinds classification tasks are `quanteda`, `Readme`, and `RTextTools` in `R`, and especially `scikit-learn` in `python`.

3.2.3 Discovery

While supervised classification is a deductive exercise in which texts are grouped into categories that are defined theoretically, other approaches are inductive and can be used to discover latent structures in the corpus and situate the texts within this latent structure.

Well-known examples in political science include "wordfish" (Slapin and Proksch, 2008; Lowe, 2008), which maps texts onto ideal points on ideological or issue-specific dimensions (Lowe, 2013). These methods need very careful text preprocessing, parameter tuning, and testing in order to be reliable.

Another strand of latent variable models are generative mixed-membership models, such as topic models, which can uncover the semantic structure of a corpus (Blei et al., 2003). Topic models can be a useful tool to identify frames in texts. As DiMaggio et al. (2013, 578, 593) write, "[m]any topics may be viewed as frames...and employed accordingly....[T]opic modeling has some decisive advantages for rendering operational the idea of 'frame'." In this context, mixed-membership means that these models assume that each document can be assigned to multiple categories, in different proportions. In other words, a given text will not include just one topic, but multiple topics, although different texts will give more or less importance to different topics. A particularly useful variant of topic models is the structural topic model (Roberts et al., 2014, 2016), which allows the prior distribution of documents and words over topics to be influenced by covariates. For instance, this allows to measure how the topics co-vary with time or with other variables of interest. For example, such models can be used to analyze how the topics of newspaper articles vary depending on whether male or female politicians are mentioned (Gilardi et al., 2018; Gilardi and Wueest, 2017). Furthermore, topic models can also be used to explore corpora in order to uncover novel measures or research questions (see Wüest, 2018).

Scaling procedures can be conducted using the packages `austin` or `quanteda` in `R`, while topic models can be estimated with `gensim` in `python` and `stm` in `R`.

3.2.4 New Directions of Text-as-Data Applications

We identify three main directions in which text-as-data applications are currently developing. First, causality. Most text-as-data approaches are purely predictive, but social scientists and policy analysis are often interested in establishing causal relationships. Causal inference with text data is not straightforward, but Egami et al. (2018) have put forward a framework to facilitate this task. Egami et al. (2018) suggest to estimate the causal effects in sequential experiments. Concretely, the advice is to split the data in a similar way as in the training of supervised classifications. Concretely, one set should be used to optimize the inductive procedures (for example, the number of topics in a topic model), while the other should be used for estimating causal relationships.

Second, another stream of research aims to go beyond the many simplifications conducted when preprocessing the texts, which usually involve discarding a lot of potentially useful information. Computer science research on word embeddings and artificial neural networks, or "deep learning" as it is often referred to, will likely gain prominence in text-as-data applications in political science and poliy analysis (Mikolov et al., 2013; LeCun et al., 2015).

Third, the "as-data" trend will not stop at text. The next frontier involves images, sounds, and videos (e.g., Dietrich et al., 2018; Joo and Steinert-Threlkeld, 2018).

4 Conclusion

The goal of this chapter was to offer an overview of text-as-data methods for comparative policy analysis, a field in which these methods have been used less extensively than in political science more generally. These methods have a very high potential to develop new tests of existing theories and to uncover new aspects of policy making that were previously very hard to study. In particular, we identify a number of advantages compared to traditional approaches relying on manual coding.

First, text-as-data approaches are scalable. That is, once the method has been developed or adapted for a specific project, additional material can be analyzed with sharply decreasing marginal costs. This is especially useful for projects that are intended to continue over a long period of time.

Second, text-as-data approaches make it easier to extend projects to new areas. Here, there are still non-trivial start-up costs to adapt existing procedure to new contexts, but the extension can be done much more efficiently than with traditional approaches. Often, new contexts involve new languages. Many text-as-data approaches already work quite well using automatic translation, whose performance is improving quickly.

Third, text-as-data approaches permit retroactive adjustments to data that were previously coded. With manual approaches, this is practically impossible because of the prohibitive costs.

Once a large dataset has been coded manually, for all intents and purposes it is fixed. But exploring alternative coding strategies can be useful for a number of reasons. First, one could question the original choices on theoretical or substantive grounds. Second, for projects that run over long periods, such as the Comparative Agendas Project, the categories do not necessarily remain constant. New issues emerge, and others lose importance. Adjusting the coding scheme retroactively, to ensure comparability over time, is much easier using automated approaches.

Fourth, automated approaches increase transparency and facilitate replication. Replicating the original coding is prohibitively costly if it was conducted manually. With text-as-data approaches, it is in principle possible to retrace how a dataset was constructed, starting with the raw data. Of course, this depends in large part on how carefully researchers document their procedures and make all data available, which might also be a problem for copyright or privacy reasons. However, the bar for replicability is definitely lower when using automated approaches.

Fifth, one might be discouraged by the fact that the methods available today are not quite advanced enough to do something that could be done using manual approaches. However, methods are improving very quickly, and one should focus more on what will be possible in a couple of years than on what can be done right now, keeping in mind the other advantages of automated approaches.

To conclude, we encourage scholars to invest time learning these methods, and the comparative policy analysis community to offer training to make the learning curve less steep. We hope that this chapter is a helpful starting point.

References

Baumgartner, F. R., Green-Pedersen, C., and Jones, B. D. (2006). Comparative studies of policy agendas. *Journal of European Public Policy*, 13(7):959–974.

Baumgartner, F. R. and Jones, B. D. (1993). *Agendas and Instability in American Politics*. University of Chicago Press, Chicago.

Baumgartner, F. R. and Jones, B. D. (2018). *U.S. Policy Agendas Project*.

Baumgartner, F. R., Jones, B. D., and Wilkerson, J. (2011). Comparative studies of policy dynamics. *Comparative Political Studies*, 44(8):947–972.

Benoit, K., Conway, D., Lauderdale, B., Laver, M., and Mikhaylov, S. (2016). Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review*, 110(2):278–295.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning*, 3:993–1022.

Boushey, G. (2016). Targeted for diffusion? How the use and acceptance of stereotypes shape the diffusion of criminal justice policy innovations in the american states. *American Political Science Review*, 110(1):198–214.

Braun, D. and Gilardi, F. (2006). Taking 'Galton's problem' seriously: Towards a theory of policy diffusion. *Journal of Theoretical Politics*, 18(3):298–322.

Card, D., Boydstun, A. E., Gross, J. H., Resnik, P., and Smith, N. A. (2015). The media frames corpus: Annotations of frames across issues.

Card, D., Gross, J., Boydstun, A., and Smith, N. A. (2016). Analyzing framing through the casts of characters in the news. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1410–1420.

Ciocan, D. and Wueest, B. (2017). *How MENA Media Frame the Arab Spring*. paper presented at the annual meeting of the American Political Science Association, USA, San Francisco.

Collingwood, L. and Wilkerson, J. (2012). Tradeoffs in accuracy and efficiency in supervised learning methods. *Journal of Information Technology & Politics*, 9(3):298–318.

Conway, D. and White, J. M. (2012). *Machine learning for Hackers. Case Studies and Algorithms to Get You Started*. O'Reilly, Cambridge, MA.

de Vries, E., Schoonvelde, M., and Schumacher, G. (2018). No longer lost in translation. Evidence that Google Translate works for comparative bag-of-words text applications. *Political Analysis*.

Denny, M. J. and Spirling, A. (2017). *Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It*. unpubl. Ms., Pennsylvania State University.

Dietrich, B. J., Enos, R. D., and Sen, M. (2018). Emotional arousal predicts voting on the us supreme court. *Political Analysis*.

DiMaggio, P., Nag, M., and Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding. *Poetics*, 41(6):570–606.

Dolowitz, D. and Marsh, D. (1996). Who learns what from whom: a review of the policy transfer literature. *Political Studies*, 44(2):343–357.

Dolowitz, D. P. (2000). Introduction to the special issue on policy transfer. *Governance*, 13(1):1–4.

Dowding, K., Hindmoor, A., and Martin, A. (2016). The comparative policy agendas project: Theory, measurement and findings. *Journal of Public Policy*, 36(1):3–25.

Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E., and Stewart, B. M. (2018). *How to Make Causal Inferences Using Texts*. unpubl. Ms., NJ, Princeton University.

Elder, C. D. and Cobb, R. W. (1984). Agenda-building and the politics of aging. *Policy Studies Journal*, 13(1):115–129.

Evans, J. A. and Aceves, P. (2016). Machine translation: mining text for social theory. *Annual Review of Sociology*, 42:21–50.

Fisher, D. R., Leifeld, P., and Iwaki, Y. (2013a). Mapping the ideological networks of american climate politics. *Climatic change*, 116(3-4):523–545.

Fisher, D. R., Waggle, J., and Leifeld, P. (2013b). Where does political polarization come from? locating polarization within the us climate change debate. *American Behavioral Scientist*, 57(1):70–92.

Flores, R. D. (2017). Do anti-immigrant laws shape public sentiment? a study of arizona's sb 1070 using twitter data. *American Journal of Sociology*, 123(2):333–384.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1.

Gerner, D. J., Schrodt, P. A., Francisco, R. A., and Weddle, J. L. (1994). Machine coding of event data using regional and international sources. *International Studies Quarterly*, 38(1):91–119.

Gilardi, F., Shipan, C. R., and Wüest, B. (2018). Policy diffusion: The issue-definition stage. University of Zurich and University of Michigan.

Gilardi, F. and Wueest, B. (2017). *Newspaper coverage of female candidates during election campaigns: Evidence from a structural topic model*. paper presented at the annual meeting of the American Political Science Association, USA, San Francisco.

Golder, S. A. and Macy, M. W. (2014). Digital footprints: Opportunities and challenges for online social research. *Annual Review of Sociology*, 40:129–152.

González-Bailón, S. (2017). *Decoding the Social World: Data Science and the Unintended Consequences of Communication*. MIT Press, Cambridge, MA.

Graham, E. R., Shipan, C. R., and Volden, C. (2013). The Diffusion of Policy Diffusion Research in Political Science. *British Journal of Political Science*, 43(3):673–701.

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. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21:267–297.

Hopkins, D. and King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247.

Jatana, N., Puri, S., Ahuja, M., Kathuria, I., and Gosain, D. (2012). A survey and comparison of relational and non-relational database. *International Journal of Engineering Research & Technology*, 1(6).

John, P. (2006). The policy agendas project: a review. *Journal of European Public Policy*, 13(7):975–986.

Jones, B. D. and Baumgartner, F. R. (2005). *The Politics of Attention. How Government Prioritizes Problems*. University of Chicago Press, Chicago.

Joo, J. and Steinert-Threlkeld, Z. C. (2018). Image as data: Automated visual content analysis for political science. *arXiv preprint arXiv:1810.01544*.

Jungherr, A. and Theocharis, Y. (2017). The empiricist's challenge: Asking meaningful questions in political science in the age of big data. *Journal of Information Technology & Politics*, 14:97–109.

Jurafsky, D. and Martin, J. H. (2009). *Speech an Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall, Upper Saddle River, NJ.

Jurka, T. P., Collingwood, L., Boydstun, A. E., Grossman, E., and van Atteveldt, W. (2013). Rtexttools: A supervised learning package for text classification. *The R Journal*, 5(1):6–12.

King, G., Pan, J., and Roberts, M. E. (2013). How censorship in china allows government criticism but silences collective expression. *American Political Science Review*, 107(2):326–343.

Klüver, H. (2013). Lobbying as a collective enterprise: winners and losers of policy formulation in the european union. *Journal of European Public Policy*, 20(1):59–76.

Klüver, H. and Mahoney, C. (2015). Measuring interest group framing strategies in public policy debates. *Journal of Public Policy*, 35(2):223–244.

Klüver, H., Mahoney, C., and Opper, M. (2015). Framing in context: how interest groups employ framing to lobby the european commission. *Journal of European Public Policy*, 22(4):481–498.

Krippendorf, K. (2004). *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, Thousand Oaks.

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

Lazer, D. and Radford, J. (2017). Data ex machina: Introduction to big data. *Annual Review of Sociology*, 43:19–39.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521:436–444.

Leifeld, P. (2013). Reconceptualizing major policy change in the advocacy coalition framework: A discourse network analysis of german pension politics. *Policy Studies Journal*, 41(1):169–198.

Leifeld, P. and Haunss, S. (2012). Political discourse networks and the conflict over software patents in europe. *European Journal of Political Research*, 51(3):382–409.

Linder, F., Desmarais, B. A., Burgess, M., and Giraudy, E. (2018). Text as policy: Measuring policy similarity through bill text reuse. *Policy Studies Journal*, page forthcoming.

Loftis, M. W. and Mortensen, P. B. (2018). Collaborating with the machines: a hybrid method for classifying policy documents. *Policy Studies Journal*, page forthcoming.

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

Lowe, W. (2013). *Putting it all on the line: Some unified theory for text scaling*. Paper prepared for the American Political Science Association meeting September 2013, IL, Chicago.

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., and Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis*.

Meyer, D. (2012). *Support Vector Machines*. Technische Universität Wien, Austria.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. CoRR, abs/1301.3781.

Monroe, B. L. and Schrodt, P. A. (2008). Introduction to the special issue: The statistical analysis of political text. *Political Analysis*, 16(4):351–355.

Müller, L. (2015). *Comparing Mass Media in Established Democracies. Patterns of Media Performance*. Palgrave Macmillan, London, UK, and New York, NY.

Nowlin, M. C. (2016). Modeling issue definitions using quantitative text analysis. *Policy Studies Journal*, 44(3):309–331.

Pierson, P. (1993). When effect becomes cause: Policy feedback and political change. *World Politics*, 45:595–628.

Pierson, P. (2000). Increasing returns, path dependence, and the study of politics. *American Political Science Review*, 94(2):251–267.

Proksch, S.-O., Lowe, W., Wäckerle, J., and Soroka, S. (2018). Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches. *Legislative Studies Quarterly*.

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

Roberts, M. E., Stewart, B. M., and Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515):988–1003.

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

Sebók, M. and Berki, T. (2017). Incrementalism and punctuated equilibrium in hungarian budgeting (1991-2013). *Journal of Public Budgeting, Accounting & Financial Management*, 29(2):151–180.

Simmons, B., Dobbin, F., and Garrett, G. (2006). Introduction: The international diffusion of liberalism. *International Organization*, 60(4):781–810.

Slapin, J. B. and Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722.

Tennant, J. P., Waldner, F., Jacques, D. C., Masuzzo, P., Collister, L. B., and Hartgerink, C. H. J. (2016). The academic, economic and societal impacts of open access: an evidence-based review. *F1000Research*, 5(632).

Walker, J. L. (1969). The diffusion of innovations among the American states. *American Political Science Review*, 63(3):880–899.

Wilkerson, J. and Casas, A. (2017). Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science*, 20:529–544.

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

Wüest, B. (2018). *Selective attention and the information environment. Citizens' perceptions of political problems in the Swiss federal election campaign 2015*. Paper proposal for the Special Issue in the Swiss Political Science Review on the 2015 Swiss National Election.

Wüest, B., Bütikofer, S., van der Lek, A., and Gantenbein, F. (2016). *Selects Media Analyses 2015. Election Campaign in Swiss National Media. Codebook & Technical Report*. University of Zurich – Selects – FORS, Zurich.

Young, L. and Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29(2):205–231.