Text-as-Data Methods for Comparative Policy Analysis

Fabrizio Gilardi[*] Bruno Wueest[†]

Work in progress
March 9, 2018

Abstract

Text-as-data methods—techniques relying on the automated or semi-automated analysis of text—have become increasingly prevalent in social science. This paper discusses their potential for comparative policy analysis, a field in which these methods have not been applied as extensively as in other areas. The paper first applications of text-as-data approaches 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. We conclude by highlighting the potential of text-as-data methods for comparative policy analysis despite their relatively sparse use so far.

[*] Department of Political Science, University of Zurich, https://www.fabriziogilardi.org/

[†] Department of Political Science, University of Zurich, wueest@ipz.uzh.ch

1 Introduction

Beyond the hype of Big Data, there is no doubt that, taken together, the internet and computational social science tools have had a very significant impact on 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, both because they have always been a primary data source for social scientists and because they have become particularly plentiful and accessible with relative ease in the internet age. 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 they can both apply existing theories to new data and uncover new phenomena that previously remained hidden (Evans and Aceves, 2016).

Text-as-data approaches are becoming mainstream in political science (Grimmer and Stewart, 2013; Lucas et al., 2015; Wilkerson and Casas, 2017). The focus of this paper is on comparative policy analysis more specifically, a field in wich text-as-data methods have not yet been applied widely despite their potential. The goal of the paper is to offer an overview of existing applications and, especially, of the options and workflow of text-as-data approaches for comparative policy analysis.

Section 2 discusses studies applying text-as-data methods in comparative policy analysis. Section 3 first outlines the overall workflow and specific steps common to all text-as-data project, and then presents three specific kinds of applications: concept identification, classification, and discovery. We conclude by highlighting the potential of text-as-data approaches for comparative policy analysis.

2 Text-as-data applications in comparative policy analysis: an overview

Applications of text-as-data methods in comparative policy analysis are relatively rare, despite their potential. For instance, the successful and influential Comparative Agendas Project has continued to rely on manual coding to classify legislation and other relevant texts into 21 major topics and 220 subtopics.[1] Recent studies relying on this approach include, for example, a large collaborative study of the link between media coverage and parliamentary agendas (Vliegenthart et al., 2016). Au-

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

tomated approaches have considerable potential for coding policy agendas, but have yet to be applied systematically (Dowding, Hindmoor and Martin, 2016, 12–13). The literature on discourse networks relies on text analysis to measure discourse coalitions quantitatively through netwirk analysis Leifeld and Haunss (2012); Leifeld (2013); Fisher, Waggle and Leifeld (2013); Fisher, Leifeld and Iwaki (2013). However, because the texts are coded qualitatively, the approach does not count as text-as-data in the narrow sense, although there is certainly potential for integrating it.

Although text-as-data applications are not mainstream in comparative policy analysis, there are a few notable exceptions.

Issue definition. Nowlin (2016) discusses how topic models can be used to study how issues are defined and applies the approach to Congressional hearings regarding used nuclear fuel. Gilardi, Shipan and Wüest (2017) 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. The Policy Frames Project uses machine learning to track media tone and framing in a variety of areas (Card et al., 2015, 2016).

Policy diffusion. In addition to identifying how the issue of smoking restrictions is defined, Gilardi, Shipan and Wüest (2017) study how issue definitions diffuse across US states, finding that practical aspects of smoking restrictions are more subject to diffusion than normative rationales. Wilkerson, Smith and Stramp (2015) use a text reuse approach to trace how ideas spread from one piece of legislation to the other. Linder et al. (2018) show how the approach can be used to measure policy similarity more in general.

Lobbying. Klüver (2013) applies Wordfish (Slapin and Proksch, 2008), a scaling method, to measure the policy preferences of interest groups based on their submissions in online consultation of the European Commission. Again using consultations, Klüver, Mahoney and Opper (2015); Klüver and Mahoney (2015) use cluster analysis to identify the frames used by interest groups as well as their determinants and effectiveness.

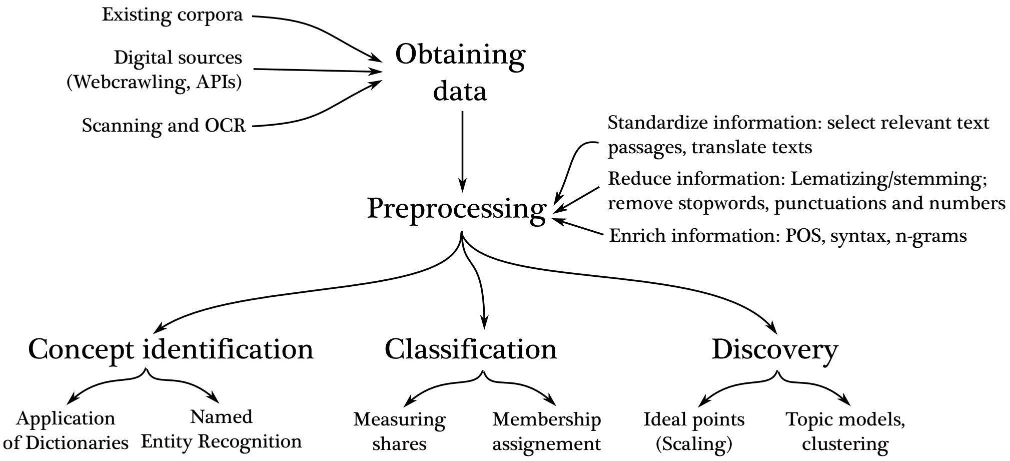
Policy feedbacks. Flores (2017) uses a dictionary approach to measure the sentiment of tweets to identify how public opinion reacts to anti-immigrant legislation, finding that the policy affected public discourse not by changing attitudes, but by mobilizing people already critical of immigrants.

3 Text-as-data in practice

3.1 Overview

The unstructured nature of text data requires research strategies that differ in many aspects from the usual quantitative approaches in policy analysis. In essence, quantitative models of text are noisier, and consequentially require more emphasis on data management, machine learning and unsupervised, exploratory analytical approaches. Figure 3.1 shows the most essential 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 be clear about how the documents should be preprocessed. 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.

Figure 1: Overview of text-as-data applications


Existing corpora
Digital sources (Webcrawling, APIs)
Scanning and OCR
Obtaining data
Standardize information: select relevant text passages, translate texts
Reduce information: Lematizing/stemming; remove stopwords, punctuations and numbers
Enrich information: POS, syntax, n-grams
Preprocessing
Concept identification
Classification
Discovery
Application of Dictionaries
Named Entity Recognition
Measuring shares
Membership assignement
Ideal points (Scaling)
Topic models, clustering

A very general topic that needs careful attention is validation. The concepts, methods and measures to secure the validity of text-as-data applications are crucial, which is why they will be discussed separately in the next section.

3.2 Machine learning

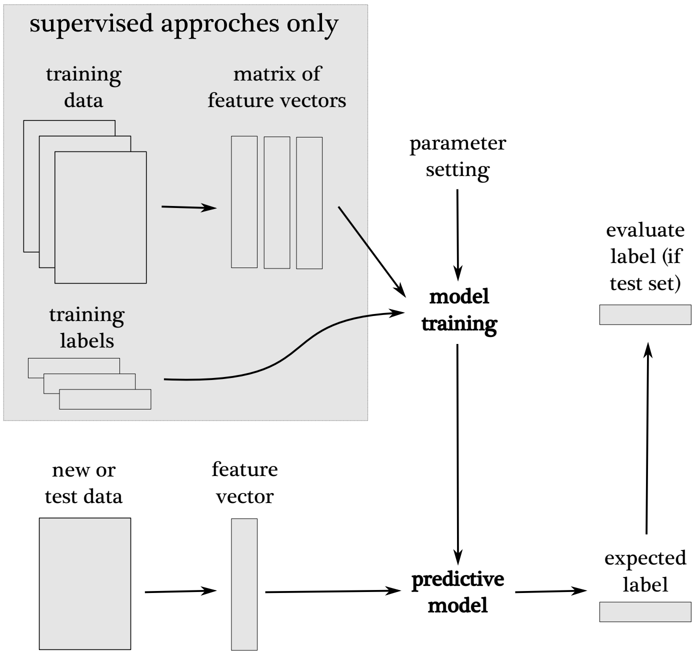
Automated text analysis is a highly insecure business. On the side of the prediction, applications of automated text analysis mainly rely on statistical and/or heuristic algorithms that retrieve information from bag-of-words representations – basically word frequency distributions – of the original texts. This means that a lot of information conveyed in text documents, from morphological information such as word order and word ambiguities to semantic information such as irony, metaphors or double negatives, are often not included in the estimations.

The implications are, on the one hand, that all results of automated models of language at best are useful approximations of the quantities of interest (Grimmer and Stewart, 2013). On the other hand, it means that automated models of language necessarily are highly domain-specific. Accordingly, there are no globally best methods for retrieving certain information from texts. The solution, of course, is what usually is termed machine learning, i.e. to focus on a careful training and validation of the algorithms used. In contrast to the probably widespread popular opinion that automation entirely replaces human judgement, human input remains key for all applications of machine learning. As for supervised approaches, researchers need to provide conceptually clear and empirically effective training labels. As for unsupervised approaches, researchers need to synthesize results and translate them into meaningful interpretations.

The general workflow of most applications of machine learning can be illustrated as shown in figure 3.2. In supervised approaches, a training corpus needs to established, which essentially means that the researchers need to manually assign labels to a sample of the documents included in the analysis. The documents and labels of this training corpus are then used for the training of the model. More specifically, the model parameters are optimized by fitting the model first to the training data, which consists of feature vectors – most commonly weighted word distributions – and the a priori known labels. Speaking in the language of inferential statistics that may be more accessible to researchers in comparative policy analysis, the documents are transformed into independent variables, i.e. the feature vectors, and the model tries to estimate their explanatory power for the dependent variable, i.e. the annotated labels.

Such a trained model can then, on the one hand, be used to predict the labels of a held-out set of the training data in order to validate the model. Such a development of a model on the basis of a comparison with manually annotated data is often also referred to as *internal validation* or out-of-sample evaluation, and the data used for comparison is usually called the *gold standard*. The biggest challenge for this validation is overfitting. Overfitting occurs if a model by change fits disproportionally well to the training data, hence the validation will overestimate the models' predictive strength. To guard oneself from overfitting, cross-validation is recommended, which means the repeated training and validation of a method on different samples of the training data. On the other hand, the model can be used to predict outcomes on previously unseen documents, which in most cases of course is the bulk of the data.

Figure 2: Stylized machine learning workflow



For unsupervised approaches, a model is directly applied to the features extracted from the text data.

While internal validation is only possible for supervised text-as-data methods, the *predictive validation* is available for both supervised and unsupervised approaches. In this case, a sample of the estimated outcome can be compared to externally compiled results. These external results can be manual annotations or automated estimations on the same data of the already existing data such as dates of policy implementations that can be used to validate results of classifications for this policy. For all options, it is crucial that the validation is conducted completely independent from the actual outcome. For example, the manual annotation of a randomly selected test sample should be performed without knowing the predictions of the model of interest.

In any case, the validation of a text-as-data application is key to ensure its reliability and replicability. Hence, the validity of the estimations should always be quantified and reported in order to meet scientific standards. The most common quantifications are based on counts of the true positives, false positives and false negatives. True positives are cases recognized correctly in both the gold standard and the held-out data set. Cases that are identified as false negatives are recognized in the gold standard but not in the held-out data set. False positives, by contrast, are recognized in the held-out data set but not in the gold standard.

From these counts, we can compute the recall (R) and precision (P) of the test data set compared to the gold standard as follows:

$R = \frac{True\,positives}{True\,positives + False\,negatives}$; $P = \frac{True\,positives}{True\,positives + False\,positives}$

Hence, the recall indicates how often a label annotated by the gold standard can also be identified by the method under concern. The precision, in contrast, indicates how often the compared method is correct when it recognizes or annotates a concept. The F1-score (F1), which is often indicated in text-as-data applications, is the harmonic mean between precision and recall[2]:

$F1 = \frac{2xRxP}{P+R}$

3.3 Workflow

3.3.1 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 legislatives 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 other 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 retrieve a large number of documents within a reasonable time frame. 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. The same holds for the fact that most original data from commercial

[2] It is recommended to additionally weight the calculation of the F1-score by the distribution of the labels in the held-out data set

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 grey 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.[3]

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.

3.3.2 Preprocessing

Given the inherently unstructured nature of text data, it needs careful preparation before it can be analyzed. This preprocessing of the text data can be presented in three steps. In the beginning of preparing a corpus, researchers need to invest in the standardization of the texts. On the one hand, several technical details such as the standardization of character encodings[4] and the extraction of meta-data (publication dates, authors etc.) need to be clarified. On the other hand, the relevant text passages, i.e. the units of observation, need to be defined. 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

[3] Here, university bodies such as the central library can provide essential support.

[4] Especially if documents stem from different sources, the encoding may vary depending on the operations systems and software programs used to process these documents. We recommend to standardize texts into one of the most common encodings that also correctly encode special characters such as German umlauts, e.g. utf8 or latin1.

discussing other topics aside.

A further crucial aspect to consider during the first step of the 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. [5] The latter, in contrast, suffers from the problem that the results produced by the different models may not be straightforwardly comparable.

The second step is the consideration of methods to the reduce the information in the texts, i.e. to filter stop words[6], punctuations and numbers. The rationale to reduce this information is that not all text elements bear important information for the word distributions used in the estimations. This also means that it will depend on the task whether the removal adds value to the estimations. Punctuations, 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 include all these methods into the machine learning workflow (see below), and, accordingly, to test empirically whether they improve the estimations. The method that generally is useful is to either stem or lemmatize the words in the texts first[7]. For most languages beside English, lemmatization should be preferred since there are many irregular conjugations and declinations.

Finally, researchers can also opt to enrich the text data by identifying the part-of-speech (POS) of words, building n-grams (combinations of words), and by 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.

The preprocessing of the text data has a decisive influence on the results of all the methods presented further below (Jurafsky and Martin, 2009). It is therefore recommended to either extensively test the influence of every step or even better

[5] Although some political scientist trust standard machine translations for this task (e.g. Lucas et al. 2015), our own experience tells us to be careful, and we at least recommend extensive qualitative tests of such translations.

[6] Though there is no single universal list of stop words used by all natural language processing tools, lists of stop words usually entail the most common words such as ‘and’ or ‘the’ in English

[7] Stemming means to cut word endings, lemmatization refers to the transformation of each word into its basic form.

to include the preprocessing parameters into the machine learning workflow (Denny and Spirling, 2017).

There are several *R* libraries and *python* packages that include one or more of the methods discussed. These are *spacyR*, *TM*, *quanteda* or *udpipe* in *R* as well as *NLTK*, *spacy* and *polyglott* in python.

3.4 Applications and methods

Text-as-data applications – or text mining or statistical natural-language processing – are one of the most fast-moving areas of political science methodology these days. As the range of applications grows rapidly, it becomes difficult to keep track of all developments. 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).

3.4.1 Concept identification

As already discussed, the goal of concept identification is to find and extract the specific text passage that refers to the concept of interest. The concept of interest thereby can be highly abstract such as the relevant political actors and institutions in a specific policy area or rather straightforward such as the names of governors of US states. In broad terms, concept identification methods can be separated into dictionary-based approaches and named entity recognition. Applications of dictionaries[8], i.e. the matching of keywords in the texts of interests, are sometimes sneered for their simplicity. However, if the operationalization leads to a comprehensive set of keywords that can unambiguously be mapped 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, but these approaches are more and more replaced by supervised classifications. Technically, the matching of dictionaries can be implemented using *Regular expressions*, which are part of the base distribution in both *R* and *python*.

A more complex set of methods to conduct concept identification is usually

[8] Depending on the literature, dictionaries are also referred to as 'ontologies', 'lists' or 'gazetteers' in the literature.

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*) 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 mostly are 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, Shipan and Wüest, 2017; Ciocan and Wueest, 2017).

3.4.2 Classification

As already outlined, 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. in which policy debate the documents of interest can be organized, 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, Pan and Roberts, 2013).

The set-up thereby is similar for every application. First, a training set needs to be build, which mostly means manually labelling a sample of the data that needs 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 established that has its parameters optimized using the labelled input and calculates probabilities for documents to labels – in the case of membership assignments – and labels to document collections – in the case of measuring shares. Popular algorithms implementing such models are multinomial naïve Bayes (Conway and White, 2012), support vector machines (Meyer, 2012), regularized paths for generalized linear models (Friedman, Hastie and Tibshirani, 2010) and maximum entropy (Jurka et al., 2013). Such models can be additionally optimized using bootstrapped training and cross-validation, evaluating the best trade-off between false positives and false negatives[9], and building ensembles, i.e. classifiers that include several algorithms or models and perform classification by a voting among them.

Furthermore, in many text-as-data projects classifications are necessary first steps in order to compile the corpus of interest, since fuzzy text data collection require a

[9] Often denoted as optimization of the receiver operator characteristic (ROC).

filter before the actual analyses (e.g. Wüest et al., 2016; Gilardi, Shipan and Wüest, 2017; Ciocan and Wueest, 2017). Software tools that allow several classification tasks are *quanteda*, *Readme* and *RTextTools* in *R* and above all *scikit-learn* in *python*.

3.4.3 Discovery

While supervised classification is a deductive exercise in which texts are grouped into theoretically defined classes, models used to discover latent structures in the corpus and place the texts within this hidden structure in principle are inductive approaches.

The most renowned approaches of models that discover latent information in political science, e.g. 'wordfish' or 'wordscores' (Slapin and Proksch, 2008; Lowe, 2008), are trying to map texts as ideal points on ideological or issue-specific dimensions (Lowe, 2013). Our own experience with these models is that they are rather unstable, i.e. that need very careful text preprocessing, parameter tuning and testing in order to be useful.

Another strand of latent variable models are generative mixed-membership models, mostly topic models using a variant of latent dirichlet allocation or non-negative factor maximization in order to uncover the semantic structure of a corpus (Blei, Ng and Jordan, 2003). Mixed-membership means that topic models are assuming a multi-dimensional structure and that every document relates to every dimension to a certain extent. A particularly useful variant is the structural topic model (Roberts et al., 2014), 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 change over time in diachronic corpora or how the words significant to a topics change according to the gender of the politicians described in the texts (Gilardi, Shipan and Wüest, 2017; Gilardi and Wueest, 2017). Topic models can also be used to explore corpora in order to uncover novel measures or research questions (see Wueest, 2018).

Scaling can be conducted using the package *austin* or *quanteda* in *R*, topic models can be run with *gensim* in *python* and *stm* in *R*.

3.4.4 New directions of text-as-data applications

We identify two directions in which text-as-data applications in policy analysis will likely develop in the near future. Only very recently, political science methodologist have begun to integrate the conceptual framework of causal inference into text-as-data models (Egami et al., 2018). While the estimation of causal effects on results of concept identification and supervised classification approaches seems straightforward, the Stable Unit Treatment Value Assumption (SUTVA) poses particular problems for unsupervised text-as-data models, since for such models, neither the outcome nor the treatment is known. In these cases, Egami et al. (2018) suggest to estimate the causal effects in sequential experiments, hence to split the data in a similar way as in the training of supervised classifications.

Another relevant shift is likely occurring at the level of the preprocessing methods as well as the algorithms used. As in the computer sciences, 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 policy analysis as well (Mikolov et al., 2013; LeCun, Bengio and Hinton, 2015).

4 Conclusion

The goal of this paper was to offer an overview of text-as-data methods for comparative policy analysis, a field in which they 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. 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.

References

Benoit, K., D. Conway, B. Lauderdale, M. Laver and S. Mikhaylov. 2016. "Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data." *American Political Science Review* 110(2):278–295.

Blei, D. M., A. Y. Ng and M. I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning* 3:993–1022.

Card, Dallas, Amber E Boydstun, Justin H Gross, Philip Resnik and Noah A Smith. 2015. "The Media Frames Corpus: Annotations of Frames Across Issues.".

Card, Dallas, Justin Gross, Amber Boydstun and Noah A Smith. 2016. Analyzing framing through the casts of characters in the news. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. pp. 1410–1420.

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

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

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

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

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

Evans, James A and Pedro Aceves. 2016. "Machine translation: mining text for social theory." *Annual Review of Sociology* 42:21–50.

Fisher, Dana R, Joseph Waggle and Philip Leifeld. 2013. "Where does political polarization come from? Locating polarization within the US climate change debate." *American Behavioral Scientist* 57(1):70–92.

Fisher, Dana R, Philip Leifeld and Yoko Iwaki. 2013. "Mapping the ideological networks of American climate politics." *Climatic change* 116(3-4):523–545.

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

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

Gilardi, Fabrizio, Charles R. Shipan and Bruno Wüest. 2017. "Policy Diffusion: The Issue-Definition Stage." University of Zurich and University of Michigan.

Golder, Scott A and Michael W Macy. 2014. "Digital footprints: Opportunities and challenges for online social research." *Annual Review of Sociology* 40:129–152.

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

Jatana, N., S. Puri, M. Ahuja, I. Kathuria and D. Gosain. 2012. "A Survey and Comparison of Relational and Non-Relational Database." *International Journal of Engineering Research & Technology* 1(6).

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

Jurka, T. P., L. Collingwood, A. E. Boydstun, E. Grossman and W. van Atteveldt. 2013. "RTextTools: A Supervised Learning Package for Text Classification." *The R Journal* 5(1):6–12.

King, G., J. Pan and M. E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107(2):326–343.

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

Klüver, Heike, Christine Mahoney and Marc Opper. 2015. "Framing in context: how interest groups employ framing to lobby the European Commission." *Journal of European Public Policy* 22(4):481–498.

Lazer, David and Jason Radford. 2017. "Data ex Machina: Introduction to Big Data." *Annual Review of Sociology* 43:19–39.

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

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

Linder, Fridolin, Bruce A. Desmarais, Matthew Burgess and Eugenia Giraudy. 2018. "Text as Policy: Measuring Policy Similarity through Bill Bill Text Reuse." *Policy Studies Journal* p. 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*. IL, Chicago: Paper prepared for the American Political Science Association meeting September 2013.

Lucas, Christopher, Richard A Nielsen, Margaret E Roberts, Brandon M Stewart, Alex Storer and Dustin Tingley. 2015. "Computer-assisted text analysis for comparative politics." *Political Analysis* .

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

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

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

Nowlin, Matthew C. 2016. "Modeling Issue Definitions Using Quantitative Text Analysis." *Policy Studies Journal* 44(3):309–331.

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

Slapin, Jonathan B. and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52(3):705–722.

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

Vliegenthart, Rens, Stefaan Walgrave, Frank R Baumgartner, Shaun Bevan, Christian Breunig, Sylvain Brouard, Laura Chaqués Bonafont, Emiliano Grossman, Will Jennings, Peter B Mortensen et al. 2016. "Do the media set the parliamentary agenda? A comparative study in seven countries." *European Journal of Political Research* 55(2):283–301.

Wilkerson, John and Andreu Casas. 2017. "Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges." *Annual Review of Political Science* 20:529–544.

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

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