Methods and Data - Text as Data in Environmental Economics and Policy

Eugenie Dugoua[a], Marion Dumas[b] and Joëlle Noailly[c,d,e*]

[a] Department of Geography and Environment, London School of Economics, United Kingdom

[b] Grantham Research Institute, London School of Economics, United Kingdom

[c] Graduate Institute Geneva, Switzerland

[d] VU Amsterdam, The Netherlands

[e] Tinbergen Institute, The Netherlands

May 23, 2022

Abstract

There is growing interest in using text as data in social science research, particularly in economics. The availability of large amounts of digitized text material such as social media posts, newspapers, firms' annual reports, and patents, combined with new computer techniques, makes it increasingly possible for researchers to use this type of information. The aim of this article is to discuss the potential of these techniques for the field of environmental economics and policy.

JEL codes: Q50; C89

Introduction

There is growing interest in using text as data in social science research, particularly in economics. The availability of large amounts of digitized text material such as social media posts, newspapers, firms' annual reports, and patents, combined with new computer techniques, makes it increasingly possible for researchers to use this type of information. The aim of this article is to discuss the potential of these techniques for the field of environmental economics and policy. While text-based methods are diffusing quickly in macroeconomics, finance, industrial organization,

and political science, the application of these techniques in environmental economics research is still in the early stages. We first provide a brief overview of text-as-data methods and programming tools. Then we present examples of empirical applications of these methods in environmental economics. We conclude with a summary and a discussion of the main challenges and future prospects for these methods.

Overview of text-as-data methods and programming tools

The goal of this section is to introduce readers to well-established text-as-data techniques, including text cleaning and methods for extracting meaning, as well as programming packages, in order to provide background and insights about the key methods and applications we discuss in the next section.[1]

Cleaning Text

Typically, the first step in the process of converting text into data is "cleaning", whereby many long strings of raw text are handled and transformed into quantitative features that can then be used in econometric models.

Table 1 illustrates the steps in the cleaning process, which include tokenizing, stemming, or lemmatizing, and removing stop words from sentences. Tokenizing consists of breaking down long strings of text into a list of "tokens" or words. Stemming and lemmatization both aim to reduce words to their semantic roots. For

[1] For a more detailed treatment of text-as-data methods, see for example, Benoit (2020) and Gentzkow et al. (2019).

example, “change” and “changing” come from the same root. The process of stemming mechanically removes the endings of words, while lemmatizing involves a more complex routine that identifies the root. Cleaning text often includes removing stop words (words that do not add much meaning to a sentence, such as “is” and “but”), with the list of stop words typically being context-dependent. The text is then often transformed into a “bags-of-words” representation, which lists all the words and the number of times they appear in the text. This can be done for single words (unigrams) or for a series of n consecutive terms (n-grams). For example, “climate change” is a bigram.

TABLE 1

Text Cleaning Steps

Step	Example
Starting point: a string of raw text	But, climate change is changing everything.
Lower casing and removing punctuation	but climate change is changing everything
Tokenizing	[but, climate, change, is, changing, everything]
Stemming	[but, climat, chang, is, chang, everyth]
Lemmatizing	[but, climate, change, be, change, everything]
Removing stop words	[climate, change, change, everything]
Bags-of-words (with unigrams)	[change: 2, climate: 1, everything: 1]
Bags-of-words (with bigrams)	[but climate: 1, climate change: 1, change be: 1, be change: 1, change everything: 1, change: 2, climate: 1, everything: 1]

Methods for Extracting Meaning

The key to working with text as data is finding ways to extract meaning from large quantities of words. With this in mind, we discuss the main approaches to extracting meaning below.

Sentiment Analysis and Other Dictionary-based Methods

Sentiment analysis is a method of quantifying the extent of negative or positive emotions (or sentiments) from written text. Sentiment analysis is considered to be a dictionary-based method because it relies on a *sentiment dictionary*; that is, a list of positive and negative emotional terms. Algorithms then count the frequency of these terms to construct sentiment scores, such as assigning +1 point for every word that expresses a positive sentiment and -1 for every word that expresses a negative sentiment (see Table 2). Such methods can be used to quantify other characteristics of the text, for example the extent to which the text mentions a particular topic or uses a particular tonality, such as uncertainty or optimism. Thus, researchers need only develop a list of terms that is relevant for the topic or issue they are examining (e.g., climate risk, biodiversity, uncertainty).

TABLE 2

Example of Dictionary-Based Method

Sentiment Dictionary		Document	Score
worst	-1	"This is the **worst** flood we've seen in years. We expect **dreadful** updates as authorities report back on casualties."	-2
dreadful	-1		
amazing	+1		

Supervised Machine Learning Models

Supervised machine learning models are text-as-data methods in which the text data has already been classified into specific topics or labels (for example, a set of Twitter messages hand-coded as relevant to air pollution). This data is then used to train an algorithm to learn from the most distinctive text features in order to assign labels (e.g., "NOx", "arsenic").[2] The algorithm can then predict a label for a set of other documents that were not used in the training. Such algorithms are increasingly used

[2] The types of algorithms used vary and may include support vector machines, naive Bayes, regression trees, or neural networks.

to identify the words that dictionary-based methods should include so that the dictionaries are tailored to the text rather than defined ex-ante.

Topic Modeling

Topic modeling is a text-as-data method that identifies the topics that are present in the text without relying on any pre-labeled data. More specifically, the algorithm examines which words tend to co-occur in the same documents, groups them into topics, and quantifies the extent to which a document mentions different topics. Table 3 illustrates the type of information that such methods provide (the numbers presented in the table are for illustrative purposes only). In this example, the algorithm grouped the words "heat", "temperature" and "rain" together under Topic 1, and the words "electricity", "power" and "photovoltaic" under a different topic (Topic 2). The algorithm also assigns a probability to each word; here, for example, Topic 1 can be interpreted as being 2% about heat, 1.8% about temperature, and so on.[3] The algorithm can also estimate the prevalence of each topic in each document. In Table 3, for example, Document 1 is 90% about Topic 2 and 10% about Topic 1.[4]

TABLE 3

Illustration of Topic Modeling Method

Topic 1		Topic 2	
Word	Probability	Word	Probability
heat	0.02	electricty	0.012
temperature	0.018	power	0.012
rain	0.015	photovoltaic	0.01

Documents	Proportions	
	Topic 1	Topic 2
Document 1	10%	90%
Document 2	40%	60%
Document 3	80%	20%

[3] The probabilities would sum to 1 in a real example.

[4] Some of the most commonly used topic modeling algorithms are called Latent Dirichlet Allocation (LDA) (Blei 2012) and structural topic modeling (Roberts et al. 2013).

Programming Tools

All of the algorithms mentioned above are available via R or Python programming packages. In R, *quanteda* is a popular package that offers a wide range of functionalities, from cleaning to analysis; *text2vec* supports many advanced models; and *stm* is particularly recommended for social science applications of topic modeling. In Python, we recommend *spacy* for cleaning text; *scikit.learn* for supervised algorithms; and *gensim* for implementation of sentiment analysis and topic modeling.[5]

Applications of text-as-data methods to environmental economic research

In this section, we present examples of text-as-data methods that have been used in environmental economics research. First, we discuss applications that measure three key aspects of the policy process: the environmental problem itself, the scientific and technological response, and the policy response. Because many social science research questions concern how individuals and organizations perceive (and respond to) environmental problems and regulations, we next discuss applications that study the beliefs, preferences, and actions of three types of actors: the general public, private actors, and governments and non-governmental organizations (NGOs).

Measuring Environmental Problems

[5] See Tables A1 and A2 in the online appendix for an overview of the literature cited in this article, organized by text-as-data method and the source of the text data, respectively.

Digital text in scientific articles, policy documents, or memos that report field observations often contains information that can be helpful for measuring and understanding environmental change. For example, Nunez-Mir et al. (2016) show how text-as-data methods can be used in the field of ecology to summarize large volumes of scientific literature on forest fragmentation. In another example, Selles et al. (2020) use a topic model to examine how the definition of forest resilience in scientific articles (which focuses on the capacity of a system to adjust and recover following a disturbance) differs from the definition used in planning documents from the US Forest Service (which focuses on the ability of ecosystems to repel non-native species).

Beyond the information contained in the text of official documents, individuals report observations about their surroundings on social media, providing a rich and low-cost source of information on environmental issues (e.g., natural disasters, traffic congestion, pollution accidents), especially when measurements may be scarce. For example, Sachdeva et al. (2017) use daily tweets on the 2014 Californian wildfires to model the spatio-temporal diffusion of smoke and air pollution. Similarly, Jiang et al. (2015) count the frequency and analyze the sentiment of geotagged posts about urban air pollution on Chinese Twitter; they argue that such textual data could be used to monitor air pollution dynamics because they correlate with the Air Quality Index published by China's Ministry of Environmental Protection.

Measuring the Scientific and Technological Response

Quantitative text methods are useful for characterizing the science, innovation, and technologies related to environmental issues. For example, when administrative and

firm-level databases contain text that describes occupations, technologies, products, or industries, such texts could potentially be used to quantify the "greenness" of these activities. Other potential applications involve using the text of patents or scientific articles. Two examples are worth highlighting in this regard.

First, Myers et al. (2021) quantify the technological and geographical spillovers created by US Department of Energy R&D grants. More specifically, they calculate the textual similarity (and thus the spillovers) between patent abstracts and the grants' research objectives. Second, Dugoua (2020) uses text-as-data methods to identify trends in the number of patents and scientific articles (un)related to molecules contained in CFC substitutes to examine the induced innovation effects of the Montreal Protocol on Substances that Deplete the Ozone Layer.

Beyond providing measures to characterize science and innovation, text-as-data methods can actually advance the state of science itself. A case in point is Munõz et al. (2019), who develop the Biodiversity Observations Miner, which is a text mining tool that identifies biodiversity observations stored in scientific literature (e.g., specific species interactions). This tool can be used to select articles for large-scale meta-analyses and to rapidly summarize observations from the literature.

Finally, quantitative text methods also allow researchers to study trends in particular disciplines. For example, Polyakov et al. (2018) quantify the evolution of research topics in the journal *Environmental and Resource Economics*. Using topic modeling, they show, for example, that conservation received more research attention after 1993 (when the Convention on Biological Diversity was enacted).

Measuring Policies

There are hundreds of thousands of pages of documents containing information about policies designed to address environmental problems. Thus, policy texts (e.g., regulations, legislation, court rulings, guidance documents, agreements, and contracts) are a key source of data for measuring policies (i.e., their scope, duration, exemptions, revisions). Despite the rapid adoption of text-as-data methods in the private sector (e.g., legal services), applications are still rare in the social sciences and even more rare in environmental economics and policy. Here, we highlight a few applications that illustrate the potential for using text as data to better characterize environmental policies.

Researchers are often interested in how specific policy features (e.g., target groups or rules) vary across time and space in order to empirically evaluate policy effectiveness. In contrast to the traditional approach of reading and hand-coding policy documents to identify these policy features, text-based methods are less time-intensive and can be replicated more easily by other researchers. For example, O'Halloran et al. (2017) use a supervised model to classify laws according to one specific institutional feature: the degree of discretion that bureaucrats have in implementing them. In another study, Ash (2016) examines 1.6 million statutes enacted by U.S. state legislatures to identify phrases that describe the tax base. Then, based on the frequency with which these phrases occur, he analyzes the causal impact of changes in the tax base on states' revenues and the effect of party control on changes in the tax code.

In a study that concerns environmental policy, Heikkila and Weible (2018) use text-as-data methods to characterize the institutional structure of oil and gas regulations in Colorado. More specifically, they compile a list of keywords that

identifies actors in the oil and gas sector and different types of governance rules among actors, such as information, authority, or enforcement rules. This allows them to identify the linkages between actors and authorities on various issues, such as air quality regulations. Noailly et al. (2022) apply a supervised model to newspaper articles and build a monthly index of US environmental policy uncertainty. They find that such uncertainty rises around election cycles and is associated with reduced venture capital funding for cleantech startups. These examples illustrate how text-as-data methods can capture a large range of policy attributes and can thus be used to improve our understanding of the impacts of environmental policy design on economic outcomes.

Measuring Public Opinion

Individuals' perceptions of environmental problems shape their beliefs and preferences for environmental policy (Millner and Ollivier 2016). Thus, it is essential to analyze how the public perceives the importance and severity of environmental issues and problems. Standard approaches for collecting information about the public's environmental preferences rely mostly on costly and time-consuming public opinion surveys. Text-as-data methods offer new opportunities for researchers to track public opinion over a range of environmental topics across various temporal and spatial scales and at relatively low cost and effort. For example, using social media posts about local temperatures, Baylis (2020) investigates preferences for weather and shows that people's emotional state (measured by 'sentiment scores' from tweets, as in Table 2) declines with cold and hot temperatures and reaches a peak at 21°C. He concludes that on average, individuals are willing to pay $5-12 to exchange a day of temperatures between 30 and 35°C for a day with temperatures

between 20 and 25°C. This suggests that text-as-data methods can provide an alternative to standard approaches for valuing willingness-to-pay.

Scholars outside the economics field have used geotagged social media posts to examine the evolution of perceptions about climate change (Cody et al. 2015; Moore et al. 2019). Because social media is used by a limited sample of the population both within and across countries, other studies have relied instead on newspapers to extract the general population's perceptions about climate change (Keller et al. 2020).

Text-as-data methods also provide opportunities to exploit text from opinion surveys. These methods make it much easier to process open-ended surveys, where people can freely write a long response, rather than relying solely on questions with pre-formulated answers (Tvinnereim et al. 2015). The main advantage of open-ended questions is that results are not pre-determined by researchers, and thus richer insights may emerge about what respondents are actually thinking. For example, using topic modeling on a two-question open-ended survey aimed at eliciting Spanish citizens' perceptions about carbon tax policies, Savin et al. (2020) identify close to 30 topics on perceptions of carbon taxes and fairness.

Measuring Attitudes and Communication from Private Actors

Text sources such as legally required communication and advertisements, press releases, and third-party monitoring offer opportunities for studying how private firms influence (and are themselves affected by) environmental change and policy.

Firms' disclosures are an excellent place to start, although some disclosures are more reliable than others. For example, the risks described in 10-K filings (annual reports filed by publicly traded companies, as required by the US Securities and

Exchange Commission) are arguably more reliable than voluntary disclosures because firms are legally required to disclose them and can be sued for inaccuracies or omissions. Thus, researchers have used such filings to measure firm-specific climate risks and to analyze how financial markets price them. Typically, the challenge is identifying the parts of the document that describe such risks. Kölbel et al. (2020) use a supervised model to classify sentences into "physical climate risks" or "low-carbon transition risks" (i.e., risks related to regulatory or technological change), with a remarkable accuracy of over 90% (i.e., 90% of sentences were correctly classified). The classifications then allow them to build a climate risk score for each 10-K filing.

When disclosures are used to measure risk exposure, disclosure statements are taken at face value. However, text-as-data methods can also shed light on the transparency and honesty of disclosures. For example, Cho et al. (2010) focus on how firms use language to manage stakeholder impressions. The authors score the 10-K filings based on their tone of voice and find that firms with worse environmental performance write environmental disclosures that are more optimistic and more evasive. Similarly, Fabrizio et al. (2019) measure the level of linguistic obfuscation in companies' voluntary answers to the Climate Disclosure Project (CDP),[6] and find that firms with poor environmental performance successfully use obfuscation to conceal their performance, which results in better ratings by the CDP.

Earnings calls can provide more reliable observations than other forms of disclosure because shareholders may scrutinize firms and challenge their disclosures. For example, Sautner et al. (2020) use earnings calls to develop firm-level measures of exposure to climate risks. They use a keyword discovery algorithm[7] developed by

[6] The CDP is an audit organization that rates firms on the basis of their climate action.
[7] The algorithm initially takes a few keywords provided by researchers and, based on those keywords, builds a more comprehensive dictionary tailored to the corpus.

King et al. (2017) to identify bigrams about physical threats, costly regulation, and technological opportunities. The authors then construct climate risk exposure variables based on the frequency of these phrases and the optimism and sense of risk in the language surrounding them.

Private actors also leave "behavioral traces" that are less prone to deliberate manipulation. For example, Reboredo et al. (2018) assess the degree of optimism in investor tweets about renewable energy, while Song et al. (2019) use measures of investor attention based on Google Trends search data. In both cases, the assumption is that tweets and internet searches truthfully reflect investors' current state of mind. The measurement of investor perception is used in a dynamic statistical model of financial stocks to reveal how investor perception affects volatility and trading volumes. Thus, these applications illustrate how variables derived from text-as-data can be used in a variety of inference strategies.

Measuring Attitudes and Communication from Governments and NGOs

Government agencies, cities, international organizations, NGOs, and activists communicate about environmental problems via text (e.g., press releases, speeches, minutes of meetings, blog posts, social media). Researchers have used these documents to study the engagement levels and motives of these actors. For instance, Boussalis et al. (2018) develop a supervised model using press releases from 82 large U.S. cities to examine why and how much cities pursue climate actions. They find that local climate vulnerability plays a more important role than political factors in explaining the variation in cities' climate discussions.

Effective and timely communication is essential for increasing the impact of climate policies. Thus, another research area focuses on the communication

strategies of governments and civil society. For example, Barkemeyer et al. (2016) examine the content of the Intergovernmental Panel on Climate Change (IPCC) Summaries for Policymakers. They find that the summaries score low on readability, with no improvement over the years, in contrast to newspapers. Muehlenbachs et al. (2011) examine communication from the U.S. Environmental Protection Agency to determine whether the fact that the agency systematically releases regulatory news on Fridays and before holidays reduces media coverage and impacts on financial markets. Bertrand et al. (2018) examine comments submitted to U.S. federal agencies during the rulemaking process. They find that the comments of non-profit groups change when they receive corporate sponsorship: their messaging becomes more similar to that of their sponsors. Finally, Farrell (2016) examines the climate skeptic movement by applying topic modeling to all press releases, policy statements, or blog articles produced by skeptic organizations and then linking them to their corporate sponsors, and finds that organizations that received corporate funding published more polarizing content on climate change.

Another strand of research seeks to infer the policy positions and ideologies of policymakers from their speeches. To do so, political scientists have developed models in which ideology is a latent variable that influences speech and can be inferred statistically from text (Lauderdale et al. 2016; Schwarz et al. 2017). These methods provide exciting opportunities for researchers to measure political disagreement on environmental policy among various public actors, as reflected in Dumas (2020), who analyses the role of variation in judges' ideologies on environmental court case decisions.

Conclusions

This article has discussed how text-as-data methods can be used to measure environmental issues, policies, knowledge, and technologies, as well the beliefs, perceptions, and strategic communication of actors concerning environmental issues. In contrast to hand-coding of text, it is now easier (and cheaper) for researchers to process large amounts of text and to do so with greater transparency and reproducibility and lower subjectivity than human coders. These new methods enable researchers to measure variables that were previously difficult to quantify, especially at scale.

Text-as-data is only beginning to be used in environmental economics. Our hope is that the overview and insights presented here will inspire researchers to expand the applications of text-as-data methods. There is a vast quantity of new data, including many sources that have not yet been explored. For example, product catalogs, job advertisements, and trade fair catalogs could be used to measure the "greenness" of activities. Planning documents, building or mining permit applications, and environmental impact assessments abound. Such data sources could be used for many applications, such as informing us about technology diffusion in developing countries. Indeed, data is often missing in developing countries, and thus print and social media sources could be used to monitor environmental trends, as well as conflicts, disasters, and migration flows. In addition, public documents (such as hearings, lobbying disclosures, and environmental and trade agreements) could be used to analyze how networks of stakeholders form and evolve over time. Similarly, text-as-data methods could be used to study the arguments for and against environmental policies as well as the persuasiveness of these arguments (e.g., from court briefs). Finally, much of the research using text-as-data has been descriptive; thus, future research is needed to expand its use in causal inference.

It is also important to recognize the challenges and limitations of text-as-data

methods. First, it can be difficult to access the relevant datasets because many archival texts have yet to be digitized. Second, researchers need more guidance on methodological standards and validation procedures. For example, there are many approaches for extracting meaning from a dataset, and researchers may find it challenging to choose and validate one approach over another. Another challenge is that computer science research is moving the technical frontier very rapidly, making it difficult for social scientists to identify the most appropriate technique. Moreover, the objectives of social scientists can be quite different from those of computer scientists; this means that additional research will likely be needed to ensure that new technical developments are consistent with social scientists' goals. Finally, it is important to recognize that textual data emerge from social communication processes in which humans are trying to influence each other. Thus, such data should not be interpreted naively, as if all of these texts were factual. Depending on the specific context and the text itself, researchers may be able to learn about the subject of the communication (e.g., tweets about wildfires), the text sender's intentions and strategy of communication (e.g., corporate actors' comments to regulatory agencies), or the effect of the communication on the recipient (e.g., the reaction of markets to climate risk disclosures). Researchers need to pay careful attention to context to determine which of these uses is appropriate and will be most effective in advancing research in environmental economics and policy.

References

Ash, E. 2016. *The Political Economy of Tax Laws in the US states.* Unpublished, University of Warwick.

Barkemeyer, R., S. Dessai, B. Monge-Sanz, B. G. Renzi, and G. Napolitano. 2016. "Linguistic analysis of IPCC summaries for policymakers and associated coverage." *Nature Climate Change* 6 (3): 311–316.

Baylis, P. 2020. "Temperature and temperament: Evidence from Twitter." *Journal of Public Economics* 184:104161.

Baylis, P., N. Obradovich, Y. Kryvasheyeu, H. Chen, L. Coviello, E. Moro, M. Cebrian, and J. H. Fowler. 2018. "Weather impacts expressed sentiment." *PloS One* 13 (4): e0195750.

Benoit, K. 2020. "Text as Data: An Overview." In *The SAGE Handbook of Research Methods in Political Science and International Relations,* edited by L. Curini and R. Franzese. SAGE Publications Ltd.

Bertrand, M., M. Bombardini, R. Fisman, B. Hackinen, and F. Trebbi. 2018. *Hall of Mirrors: Corporate Philanthropy and Strategic Advocacy.* Technical report w25329. National Bureau of Economic Research.

Blei, David M. 2012. "Probabilistic Topic Models." Communications of the ACM 55 (4): 77–84.

Boussalis, C., T. Coan, and M. Holman. 2018. "Climate Change Communication from Cities in the USA." *Climatic Change* 149 (2): 173–187.

Cho, C. H., R. W. Roberts, and D. M. Patten. 2010. "The language of US corporate environmental disclosure." *Accounting, Organizations and Society* 35 (4): 431–443.

Cody, E. M., A. J. Reagan, L. Mitchell, P. S. Dodds, and C. M. Danforth. 2015. "Climate Change Sentiment on Twitter: An Unsolicited Public Opinion Poll." *PloS One* 10 (8): e0136092.

Dugoua, E. 2020. *Induced Innovation and International Environmental Agreements: Evidence from the Ozone Regime.* Grantham Research Institute Working Papers (363).

Dumas, M. 2020. "Detecting Ideology in Judicial Language." In *Law as Data,* edited by M. Livermore and D. Rockmore. SFI Press.

Fabrizio, K. R., and E.-H. Kim. 2019. "Reluctant Disclosure and Transparency: Evidence from Environmental Disclosures." *Organization Science* 30 (6): 1207–1231.

Farrell, J. 2016. "Corporate Funding and Ideological Polarization about Climate Change." *Proceedings of the National Academy of Sciences* 113 (1): 92–97.

Gentzkow, M., B. T. Kelly, and M. Taddy. 2019. "Text as Data." *Journal of Economic Literature* 57 (3): 535–574.

Heikkila, T., and C. M. Weible. 2018. "A Semi-automated Approach to Analyzing Polycentricity." *Environmental Policy and Governance* 28 (4): 308–318.

Jiang, W., Y. Wang, M.-H. Tsou, and X. Fu. 2015. "Using Social Media to Detect Outdoor Air Pollution and Monitor Air Quality Index (AQI): A Geo-Targeted Spatiotemporal Analysis Framework with Sina Weibo (Chinese Twitter)." *PloS One* 10 (10): e0141185.

Keller, T. R., V. Hase, J. Thaker, D. Mahl, and M. S. Schäfer. 2020. "News Media Coverage of Climate Change in India 1997–2016: Using Automated Content Analysis to Assess Themes and Topics." *Environmental Communication* 14 (2): 219–235.

King, G., P. Lam, and M. E. Roberts. 2017. "Computer-assisted keyword and document set discovery from unstructured text." *American Journal of Political Science* 61 (4): 971–988.

Kölbel, J. F., M. Leippold, J. Rillaerts, and Q. Wang. 2020. *Ask BERT: How Regulatory Disclosure of Transition and Physical Climate Risks affects the CDS Term Structure.* Unpublished.

Lauderdale, B. E., and A. Herzog. 2016. "Measuring Political Positions from Legislative Speech." *Political Analysis* 24 (3): 374–394.

Millner, Antony, and Hélène Ollivier. 2016. "Beliefs, Politics, and Environmental Policy." *Review of Environmental Economics and Policy* 10 (2): 226–44

Moore, F. C., N. Obradovich, F. Lehner, and P. Baylis. 2019. "Rapidly Declining Remarkability of Temperature Anomalies may Obscure Public Perception of Climate Change." *Proceedings of the National Academy of Sciences* 116 (11): 4905–4910.

Muehlenbachs, L., E. Newcomb Sinha, and N. R. Sinha. 2011. *Strategic Release of News at the EPA.* Resources for the Future Discussion Paper 11-45.

Munõz, G., W. D. Kissling, and E. E. van Loon. 2019. "Biodiversity Observations Miner: A Web Application to Unlock Primary Biodiversity Data from Published Literature." *Biodiversity Data Journal,* no. 7, e28737.

Myers, K., and L. Lanahan. 2021. "Research Subsidy Spillovers, Two Ways." SSRN Working Paper.

Noailly, J., L. Nowzohour, and M. van den Heuvel. 2022. "A News-based Index of Environmental Policy Uncertainty." Unpublished.

Nunez-Mir, G. C., B. V. Iannone III, B. C. Pijanowski, N. Kong, and S. Fei. 2016. "Automated Content Analysis: Addressing the Big Literature Challenge in Ecology and Evolution." *Methods in Ecology and Evolution* 7 (11): 1262–1272.

O'Halloran, S., M. Dumas, S. Maskey, G. McAllister, and D. K. Park. 2017. "Computational Data Sciences and the Regulation of Banking and Financial Services." In *From Social Data Mining and Analysis to Prediction and Community Detection,* edited by M. Kaya, Ö. Erdoğan, and J. Rokne, 179–209. Cham: Springer International Publishing.

Polyakov, M., M. Chalak, M. S. Iftekhar, R. Pandit, S. Tapsuwan, F. Zhang, and C. Ma. 2018. "Authorship, Collaboration, Topics, and Research Gaps in Environmental and Resource Economics 1991–2015." *Environmental & Resource Economics* 71 (1): 217–239.

Reboredo, J. C., and A. Ugolini. 2018. “The Impact of Twitter Sentiment on Renewable Energy Stocks.” *Energy Economics* 76:153–169.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, and Edoardo M. Airoldi. 2013. “The Structural Topic Model and Applied Social Science.” In Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation.

Sachdeva, S., S. McCaffrey, and D. Locke. 2017. “Social Media Approaches to Modeling Wildfire Smoke Dispersion: Spatio-temporal and Social Scientific Investigations.” *Information, Communication and Society* 20 (8): 1146–1161.

Sautner, Z., L. van Lent, G. Vilkov, and R. Zhang. 2020. *Firm-level Climate Change Exposure.* SSRN Working Paper.

Savin, I., S. Drews, S. Maestre-Andrés, and J. van den Bergh. 2020. “Public Views on Carbon Taxation and its Fairness: A Computational-linguistics Analysis.” *Climatic Change* 162 (4): 2107–2138.

Schwarz, D., D. Traber, and K. Benoit. 2017. “Estimating Intra-Party Preferences: Comparing Speeches to Votes.” *Political Science Research and Methods* 5 (2): 379–396.

Selles, O. A., and A. R. Rissman. 2020. “Content Analysis of Resilience in Forest Fire Science and Management.” *Land Use Policy* 94:104483.

Song, Y., Q. Ji, Y.-J. Du, and J.-B. Geng. 2019. “The dynamic dependence of fossil energy, investor sentiment and renewable energy stock markets.” *Energy Economics* 84:104564.

Tvinnereim, E., and K. Fløttum. 2015. “Explaining topic prevalence in answers to open- ended survey questions about climate change.” *Nature Climate Change* 5 (8): 744– 747.

TABLE A1
Text-as-Data Methods Used in the Literature

Methods	Literature
Dictionary-based methods	
Sentiment analysis	Barkemeyer et al. (2016), Cody et al. (2015), Jiang et al. (2015), Moore et al. (2019), Reboredo et al. (2018), Sautner et al. (2020)
Other dictionaries	Ash (2016), Heikkila et al. (2018), Muehlenbachs et al. (2011), Munõz et al. (2019), Sautner et al. (2020), Cho et al. (2010)
Supervised machine learning models	
Naive Bayes classifier	O'Halloran et al. (2017)
Support Vector Machine	Boussalis et al. (2018), Noailly et al. (2021)
Deep Neural Network (BERT)	Kölbel et al. (2020)
Unsupervised machine learning models	
Latent Dirichlet Allocation	Boussalis et al. (2018), Dugoua (2020), Keller et al. (2020), Noailly et al. (2021), O'Halloran et al. (2017), Polyakov et al. (2018)
Structural Topic Modeling	Farrell (2016), O'Halloran et al. (2017), Sachdeva et al. (2017), Savin et al. (2020), Selles et al. (2020), Tvinnereim et al. (2015)

TABLE A2
Sources of Data used in the Literature

Data sources	Literature
Scientific literature	Munõz et al. (2019), Nunez-Mir et al. (2016), Polyakov et al. (2018)
Patents	Dugoua (2020), Myers et al. (2021)
Newspapers	Keller et al. (2020), Noailly et al. (2021)
Press releases	Boussalis et al. (2018), Muehlenbachs et al. (2011)
Twitter	Baylis et al. (2018), Moore et al. (2019), Reboredo et al. (2018), Sachdeva et al. (2017)
Web-scraping[8] of websites	Farrell (2016)
Policy documents	Barkemeyer et al. (2016), Selles et al. (2020)
Legislative texts	Ash (2016), Heikkila et al. (2018), O'Halloran et al. (2017), Bertrand et al. (2018) (public comments on regulatory text)
Political speeches	Lauderdale et al. (2016), Schwarz et al (2017)
10-k corporate annual reports	Cho et al. (2010), Kölbel et al. (2020)
Transcripts of earnings conference calls	Sautner et al. (2020)
Disclosure surveys	Fabrizio et al. (2019)
Citizen surveys	Savin et al. (2020), Tvinnereim et al. (2015)

[8] Tools to automatically extract data from websites