Causal inference from text: A commentary

Dhanya Sridhar[1] and David M. Blei[2]*

Statistical and machine learning methods help social scientists and other researchers make causal inferences from texts.

The science of causality is about how to make an inference about a hypothetical intervention (*1*–*4*). Suppose we are clinicians, deciding whether or not to administer a new drug to help a patient recover from a disease. To make the decision, we ask: What is the difference in recovery rates if we intervene and give patients the drug versus if we intervene and prevent them from taking it?

The field of causal inference is about how to analyze data, e.g., about whether patients take the drug and whether they recover, to answer such questions. What must we assume about data collection to be able to estimate a causal effect? Provided these assumptions hold, what is an appropriate method to analyze the data? The fundamental challenge to causal inference is that we can never observe what happens when a patient takes and does not take the drug. This crucial fact distinguishes causal inference from traditional statistics.

CAUSAL INFERENCE FROM TEXT DATA

In a causal analysis, the treatment and outcome are usually simple variables, like whether a patient takes a drug and whether that patient recovers. But in the social sciences, many causal inference problems involve language, which is a considerably more complex type of variable. How does the content of a candidate's speech affect the outcome of the election? How does showing someone an ad affect how they write about a product? Analyzing text data to answer these types of causal questions is the problem addressed in the excellent article of Egami *et al.* (*5*).

In this issue of *Science Advances*, the authors formally articulate what it means to make a causal inference from text. A text can be a "treatment," such as a candidate's speech, or it can be an "outcome," such as an open-ended survey response. [A third situation, where text confounds the causal inference, is the subject of some of our own work (*6*).]

The main wrinkle, however, is that most analyses of text data involve coding, a way of summarizing a document into a simpler set of labels, properties, or topics. The code might be manually produced as part of the data collection, such as by asking analysts to read the speeches and record the ideology that each communicates. Or the code might be automatically generated using machine learning methods. For example, the data analyst might fit a topic model (*7*) to uncover the latent themes that pervade the texts and then to use the fitted model to represent each text by the themes that it presents.

In considering this wrinkle, one of the key insights of Egami *et al.* is that when we analyze text data in a causal situation, the inference of interest is often one about the underlying coding of the text. How does the ideology communicated in the speech affect the election results? How does seeing an advertisement affect the topics discussed in the survey response?

Thus, Egami *et al.* carefully consider how the construction of the code interacts with the assumptions that underpin the causal inference. Their main takeaway: Regardless of how you code your texts—whether by hand or with algorithms—it is important to use data splitting. To protect against fundamental issues in the downstream causal estimates, you should construct your coding function with a different subset of the data than the one used to estimate causal effects.

THINKING MORE ABOUT THE CODE

Beyond this important criterion, what are other considerations when choosing a code? What should we look for in a textual code that is used for causal inference? Inspired by the framework developed by Egami *et al.* (*5*), we look further at the different relationships that a code can have with the text. These ways of thinking could lead to further research and methods for causal inference from text.

In the simplest situation, the text is the outcome. A treatment causes a text, and the text is correlated with (or causes) the code. Here, a causal inference about the effect of a treatment on the code is meaningful, regardless of the coding. That said, not all codings are equal—some codes may be affected by the treatment, and others may not. One interesting research question is how to find a coding such that there is an observable causal effect.

In the example from Egami *et al.* (*5*), the data are from an open-ended survey about immigration. The treatment involves information about an immigrant's criminal history; the response is a textual description of what the respondent thinks should happen to the immigrant. We are interested in how different treatments (histories) affect certain aspects of the survey responses. A good code is one that captures these aspects.

In other situations, the text is the treatment, a possible cause of the outcome. When the text is the treatment, there are two ways to involve the code (Fig. 1).

First, the code can capture the aspects of the text that cause the outcome, as illustrated in Fig. 1A. As one example, loan applicants write a statement when applying for a loan, and we record whether they receive or do not receive the funds. The code can represent all the information in the application that a loan officer relies on to approve or reject the application. [An associational analysis of related data, which uses topic models, is in (*8*).] In the example from Egami *et al.*, a political candidate's biography affects

[1] MILA and University of Montreal, Montréal, QC, Canada.
[2] Columbia University, New York, NY, USA.
*Corresponding author. Email: david.blei@columbia.edu

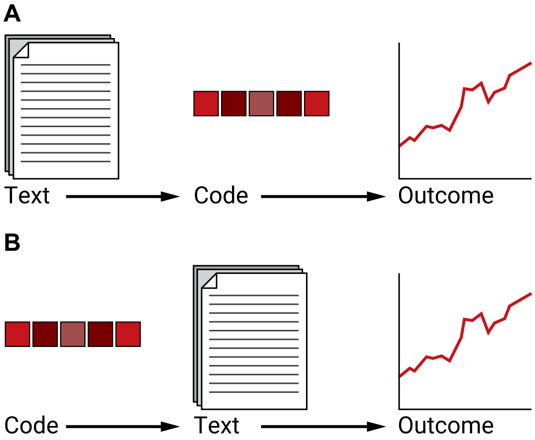

Fig. 1. Two ways that the code plays a role in a causal model that involves text. (**A**) Code captures aspects of the text that causes the outcome or (**B**) code represents aspects that affect what is written in the text. Credit: Austin Fisher, *Science Advances*.

their popularity. The coding function reduces the biography into features that causally connect to whether voters support the candidate.

When the code mediates the text and the outcome, causal inference can reveal what should be encoded in a biography to receive voter support or in a loan application to receive a loan. A good code is one that captures aspects of text that plausibly affect the outcome. Understanding such codes—how to find them and what we must assume about them—is discussed in (*9*).

In a second paradigm for text as treatment, the code directly affects how the text is written, and the text causally affects the outcome (Fig. 1B). For example, a candidate has a political ideology. It affects the language of the candidate's speeches, and the speeches affect whether the candidate is elected. A causal inference can suggest to candidates how to adapt their ideological positions to affect voting behavior.

When the code causes the text, a good coding will infer the ideology a candidate had in mind from the content of their speeches. In this sense, the code is "manipulable" (e.g., in that a candidate can choose their ideology), before producing the text. How to form such a coding, even manually, is a difficult problem because the ideal code captures the intentions of the authors of the texts.

What if we want to infer the code algorithmically, for example, with topic models? Unsupervised machine learning methods, while they often can produce interpretable representations, will not necessarily recover "manipulable" codes like these. How to guide unsupervised methods to extract such causal variables connects to the new field of causal representation learning (*10*). Developing causal representation learning for language data is a promising avenue of research.

IN SUM

The work by Egami *et al.* (*5*) advances the field of text analysis and causality by showing how to estimate causal quantities from text data, where the text is either an outcome or a treatment. The methods developed in their paper—and in the growing literature on causal inference from text—can help social scientists and other researchers make new types of causal inferences from text data.

REFERENCES

1. J. Pearl, *Causality* (Cambridge Univ. Press, ed. 2, 2009).
2. G. Imbens, D. Rubin, *Causal Inference in Statistics, Social and Biomedical Sciences: An Introduction* (Cambridge Univ. Press, 2015).
3. S. Morgan, C. Winship, *Counterfactuals and Causal Inference* (Cambridge University Press, ed. 2, 2015).
4. M. Hernan, J. Robins, *Causal Inference: What If?* (Chapman & Hall/CRC, 2020).
5. N. Egami, C. Fong, J. Grimmer, M. Roberts, B. Stewart, How to make causal inferences using texts. *Sci. Adv.* **8**, eabg2652 (2022).
6. V. Veitch, D. Sridhar, D. Blei, *Uncertainty in Artificial Intelligence* (Proceedings of Machine Learning Research, 2020).
7. M. E. Roberts, B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, D. G. Rand, Structural topic models for open-ended survey responses. *Am. J. Polit. Sci.* **58**, 1064–1082 (2014).
8. O. Netzer, A. Lemaire, M. Herzenstein, When words sweat: Identifying signals for loan default in the text of loan applications. *J. Market. Res.* **56**, 960–980 (2019).
9. C. Fong, J. Grimmer, Causal inference with latent treatments. *Am. J. Polit. Sci.* 10.1111/ajps.12649 (2022).
10. B. Schölkopf, F. Locatello, S. Bauer, N. Ke, N. Kalchbrenner, A. Goyal, Y. Bengio, Towards causal representation learning. arXiv:2102.11107 (2021).

10.1126/sciadv.ade6585

Science Advances

Causal inference from text: A commentary

Dhanya Sridhar and David M. Blei

Sci. Adv. **8** (42), eade6585. DOI: 10.1126/sciadv.ade6585

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.