Symposium: Commentaries and Rejoinder

Sociological Methodology
1–5

http://sm.sagepub.com

COMMENT: TO MEASURE MEANING IN BIG DATA, DON'T GIVE ME A MAP, GIVE ME TRANSPARENCY AND REPRODUCIBILITY

Laura K. Nelson*

*Northeastern University, Boston, MA, USA
Corresponding Author: Laura K. Nelson, l.nelson@northeastern.edu
DOI: 10.1177/0081175019863783

1. INTRODUCTION

In methodological discussions about measuring meaning in big data, one prominent opinion is that computational methods can and should replace all subjective (read: bad) decisions with statistical (read: good) procedures. From my reading of "Analyzing Meaning in Big Data: Performing a Map Analysis Using Grammatical Parsing and Topic Modeling," Goldenstein and Poschmann (hereafter GP) hold this belief. Aspiring for statistical purity belies an ontological truth about text as data: Text conveys a vast amount of information, much of it ambiguous and only some of which is relevant for a research question or purpose. As a result, sociologists using text as data must make a dizzying number of decisions about what information to extract and how to answer their research question. It is simply impossible to represent text absent any subjective decisions and have those representations be analytically useful, particularly when measuring meaning in text. GP's analysis does not escape this reality.

In this comment, I review the many subjective decisions embedded in GP's article to reframe the strength of computational methods in terms of transparency and replicability, not statistical objectivity. In an attempt

to reorient the field toward a new standard for measuring meaning in big data, one based on transparency and replicability, I propose five guidelines to evaluate any text-analysis project.

2. STATISTICAL PURITY: A FALSE GOD

GP contribute an important tool to the existing (and growing) sociological text-analysis tool kit: operationalizing *frames* by combining grammatical parsing to capture communication structure, with topic models to capture semantic context. Their article exemplifies how to expand our text-analysis tool kit by applying computational tools to address an important sociological question. Others should follow their lead.

The way GP frame their contribution to text-analysis methods, however, undermines their actual contribution to the field and obscures the true strengths of computational methods. Following Lee and Martin (2015), GP argue that text-mining tools should ideally supplant coding procedures that involve the "necessarily subjectively driven exclusion of linguistic units, or the grouping of particularities into labeled categories beyond the observer's sight" (p. XX). Their map approach, they claim, uncovers patterns of meaning without being influenced by coding procedures that require these ex ante subjective decisions, enabling "first and undistorted ex-post interpretations" (p. XX).

Far from removing ex ante choices, claims of statistical objectivity serve only to mask the subjective decisions that are necessary to measure meaning in big data. To illustrate, GP's own "undistorted" maps actually rely, in my count, on no fewer than 26 ex ante choices,[1] including subjective decisions to "[exclude] linguistic units" and decisions to group "particularities into labeled categories beyond the observer's sight" (p. XX). Each of their choices affects both the look of their maps and the meanings conveyed. In addition to broad decisions such as the way they operationalize communication structure (grammatical parsing) and semantic context (topic models) and their choice of particular software, algorithms, and measures, the authors chose to

- constrain the "semantic surrounding" to the paragraph in which their chosen key words occurred;
- include only adjectives and nouns (and exclude proper nouns) in the text used to construct their topic model;

- exclude a full 38 of the 70 semantic patterns they estimated and pool the resulting 32 topics into six semantic groups;
- use the number of unique semantic triplets (rather than frequency) per main era (era defined through yet another ex ante choice of clustering cutoff) as the relevant textual characteristic of their data;
- label the six semantic groups with their own subjectively chosen phrases.

Listing these choice points is not a criticism of their analysis. On the contrary, the fact that I could easily list every ex ante choice they made illustrates the very power of these techniques. The strength of computational text-analysis methods lies not in representing text free of ex ante interpretations but in making these necessary decisions as transparent and reproducible as possible. Claiming text-mining techniques do the former obscures their strength in the latter.

3. TRANSPARENCY AND REPLICABILITY: A BETTER GOD

As the impact of computational methods on text analysis grows and available techniques proliferate, we must be more precise about how to use these methods in sociological research. Rather than strive for statistical purity, I argue we should work toward transparency and replicability. Toward this goal, I recommend asking the following questions of any text-analysis project, regardless of method used:

1. Is the information extracted from the text the most relevant information to the social process/concept/question, and is the relevance of that information rooted in existing linguistic and sociological theory? Was there any relevant information the authors failed to extract that could challenge their conclusion?[2]
2. Were the techniques (computational or otherwise) used to extract this information the most accurate techniques available?[3]
3. Is the method used the most transparent and replicable available? If others followed the steps the authors took, would they extract the same information from the text?
4. Within reason, if the authors altered linguistic features/algorithms/other decision points, would they extract the same information from the text, or would these changes alter their conclusion in any substantial way?

5. Is the authors' interpretation reproducible? If others were to independently analyze the same data using the same methods, would they reach a similar conclusion about the social world?

GP meet four of these five proposed guidelines. Like most text-analysis projects in the social sciences, GP did few sensitivity checks (Point 4 as previously described); it is therefore difficult to know the full impact of their 26 subjective decisions on their substantive conclusions (see e.g., Denny and Spirling 2018). But their overall approach is exemplary. GP appropriately root their operationalization of corporate responsibility frames in existing sociological and linguistic theories (p. XX), perform multiple validity checks (pp. XX, XX), and provide software to replicate their analysis.

My critique of their methodological framing, however, remains. Inaccurate and irresponsible claims that algorithms and statistics are somehow completely free of human influence has led to dangerous outcomes in the world outside academia (Eubanks 2018; Noble 2018; O'Neil 2016). A similar misrepresentation of the role of computational methods in measuring meaning will inevitably lead to overstated, biased, or simply wrong conclusions in academic research. Instead, we should use these methods to make measurement transparent and replicable, to make our interpretations reproducible.

Notes

1. Not all of which I list here.
2. If, for example, you want to identify whether a specific concept (e.g., inequality) is discussed in a text, identifying all the broad themes across a corpus (e.g., by using topic models) is not the most relevant information to extract, and it will likely undercount the concept of interest (Nelson et al. 2018).
3. For example, even if topic models do reliably uncover a topic related to a particular concept of interest, other techniques, such as supervised machine learning, are likely much more accurate (Nelson et al. 2018). Researchers must know a wide range of techniques to identify the most accurate one for their purpose. Sometimes, the best technique will be qualitative, but researchers should balance this with the desire for transparency and replicability.

References

Denny, Matthew J., and Arthur Spirling. 2018. "Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It." *Political Analysis* 26(2):168–89.

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Lee, Monica, and John Levi Martin. 2015. "Coding, Counting and Cultural Cartography." *American Journal of Cultural Sociology* 3(1):1–33.

Nelson, Laura K., Derek Burk, Marcel Knudsen, and Leslie McCall. 2018. "The Future of Coding: A Comparison of Hand-Coding and Three Types of Computer-Assisted Text Analysis Methods." *Sociological Methods & Research*. doi:10.1177/0049124118769114

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.

Author Biography

Laura K. Nelson is an assistant professor of sociology in the College of Social Sciences and Humanities at Northeastern University, where she is also core faculty at the NULab for Texts, Maps, and Networks. She is affiliated faculty at the Network Science Institute and is on the Executive Committee for the Women's, Gender, and Sexuality Studies program. Previously, she has been a postdoctoral research fellow at Digital Humanities @ Berkeley and the Berkeley Institute for Data Science at the University of California-Berkeley (where she got her PhD) and in the Management and Organizations Department in the Kellogg School of Management at Northwestern University, where she was also a research affiliate at the Northwestern Institute on Complex Systems. She uses computational tools, principally automated text analysis and network analysis, to study social movements, culture, gender, institutions, and organizations. She has published in *Sociological Methods and Research* and with Oxford University Press and Springer, among other outlets, and has given invited talks and workshops on computational methods throughout the United States and internationally.