The "Trap of Metaphor": Applications and Boundaries of Topic Models in the Social Sciences

Keywords: topic models, text analysis, ontology, big data, content analysis, discourse analysis

Extended Abstract

Text analysis has been an important component of social science for a long time. Social science researchers have made numerous attempts and contributions to understanding society through various analytical methods. In recent years, with the development of the "computational turn" in social science, text analysis methods have made significant progress (Evans & Aceves, 2016; Grimmer & Stewart, 2013; Isoaho et al., 2021). Among these, statistical algorithms that analyze the latent semantic structure of documents through unsupervised learning, represented by topic modeling, have made remarkable progress in social science (Grimmer, 2010; Roberts et al., 2013).

Topic modeling, also known as probabilistic topic modeling, aims to mine the latent semantic structure behind large-scale text through probabilistic models. In such models, "topics" are usually represented by probability distributions of a series of words, and text can be represented by the probability distribution of topics. Topic modeling infers the latent topics from known text and words in reverse (Blei, 2012; Blei et al., 2003).

Since its introduction to the social science field around 2010, topic modeling has received widespread attention and application in sociology, economics, political science, management, and international relations research as a means of automated exploration of hidden information in text and revealing conceptual relationships (DiMaggio et al., 2013; Grimmer, 2010; Hannigan et al., 2019; Horowitz et al., 2019). Yet, while the majority of previous works have concentrated on enhancing the application of topic modeling from a methodological standpoint, there are relatively few articles that critically reflect on topic modeling from corpus construction to result interpretation (Isoaho et al., 2021; Nikolenko et al., 2017; Roberts et al., 2013). Existing methodological publications on topic modeling have encouraged more academics to employ the approach; yet, scholars may use it improperly due to a lack of awareness of method boundaries, even when guided by standard operating procedures. Such misuse under standardized operating procedures has raised concerns in regression analysis (Lal et al., 2021; Xu, 2022), but attention should also be paid to novel methods, particularly when scholars suggest the versatility of topic modeling and its potential to replace conventional methods in the era of big data (Isoaho et al., 2021; Jaworska & Nanda, 2018).

This article is an attempt at a critical reflection on topic modeling. By replicating several social science studies using topic modeling, we find it could be a problem to understand and characterize phenomena by labeling high-frequency co-occurring words as "topics" due to essentially metaphorical nature of “topics” (Black, 1979). Specifically, we find scholars from different backgrounds or with various research questions treat things in different nature as topics. The major problem is caused by their belief of the comparability between topic modeling and methods from their disciplines. In this study, we find even there are comparable aspects between methods, their outputs are just overlapped rather the same. In this study, we claim it as “trap of metaphor” implying that fact that the inventors of topic models used "topics"

to describe or metaphorically represent an important feature of the method, but users of the method are likely to mistakenly believe that the results of topic modeling have all the features of "topics" in everyday discourse. However, it is not our purpose to claim computational social science cannot serve for conventional theory-based studies as we find the problem can be addressed by proper selection of methods based on researchers' understanding of text. To illustrate our idea, we propose two alternatives of computational social science method that assist traditional methods to deal with big data.

References

Black, Max. 1979. "More About Metaphor." *Metaphor and thought* 2:19-41.

Blei, David M, Andrew Y Ng and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of machine Learning research* 3(Jan):993-1022.

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

DiMaggio, Paul, Manish Nag and David Blei. 2013. "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of Us Government Arts Funding." *Poetics* 41(6):570-606.

Evans, James A and Pedro Aceves. 2016. "Machine Translation: Mining Text for Social Theory." *Annual Review of Sociology* 42:21-50.

Grimmer, Justin. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political analysis* 18(1):1-35.

Hannigan, Timothy R, Richard FJ Haans, Keyvan Vakili, Hovig Tchalian, Vern L Glaser, Milo Shaoqing Wang, Sarah Kaplan and P Devereaux Jennings. 2019. "Topic Modeling in Management Research: Rendering New Theory from Textual Data." *Academy of Management Annals* 13(2):586-632.

Horowitz, Michael, Brandon M Stewart, Dustin Tingley, Michael Bishop, Laura Resnick Samotin, Margaret Roberts, Welton Chang, Barbara Mellers and Philip Tetlock. 2019. "What Makes Foreign Policy Teams Tick: Explaining Variation in Group Performance at Geopolitical Forecasting." *The Journal of Politics* 81(4):1388-404.

Isoaho, Karoliina, Daria Gritsenko and Eetu Mäkelä. 2021. "Topic Modeling and Text Analysis for Qualitative Policy Research." *Policy Studies Journal* 49(1):300-24.

Jaworska, Sylvia and Anupam Nanda. 2018. "Doing Well by Talking Good: A Topic Modelling-Assisted Discourse Study of Corporate Social Responsibility." *Applied Linguistics* 39(3):373-99.

Lal, Apoorva, Mackenzie William Lockhart, Yiqing Xu and Ziwen Zu. 2021. "How Much Should We Trust Instrumental Variable Estimates in Political Science? Practical Advice Based on over 60 Replicated Studies." *Practical Advice based on Over* 60.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley and Edoardo M Airoldi. 2013. "The Structural Topic Model and Applied Social Science." Pp. 1-20 in *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*, Vol. 4: Harrahs and Harveys, Lake Tahoe.

Xu, Yiqing. 2022. "Causal Inference with Time-Series Cross-Sectional Data: A Reflection." *Available at SSRN 3979613*.