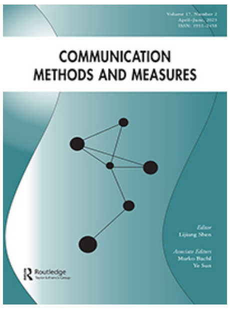

Communication Methods and Measures

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hcms20

What We Can Do and Cannot Do with Topic Modeling: A Systematic Review

Yingying Chen, Zhao Peng, Sei-Hill Kim & Chang Won Choi

To cite this article: Yingying Chen, Zhao Peng, Sei-Hill Kim & Chang Won Choi (2023) What We Can Do and Cannot Do with Topic Modeling: A Systematic Review, Communication Methods and Measures, 17:2, 111-130, DOI: 10.1080/19312458.2023.2167965

To link to this article: https://doi.org/10.1080/19312458.2023.2167965

Published online: 19 Jan 2023.

COMMUNICATION METHODS AND MEASURES
2023, VOL. 17, NO. 2, 111–130
https://doi.org/10.1080/19312458.2023.2167965

What We Can Do and Cannot Do with Topic Modeling: A Systematic Review

Yingying Chen[a], Zhao Peng[b], Sei-Hill Kim[c], and Chang Won Choi[d]

[a]School of Journalism and Communication, Renmin University of China, Beijing, China; [b]School of Communication, Emerson College, Massachusetts, USA; [c]School of Journalism and Mass Communications, University of South Carolina, South Carolina, USA; [d]School of Journalism and New Media, University of Mississippi, Mississippi, USA

ABSTRACT

Topic modeling has become an effective tool for communication scholars to explore large amounts of texts. However, empirical studies applying topic modeling often face the critical question of making meaningful theoretical contributions. In this study, we highlighted the importance of theoretical underpinning, the research design, and the methodological details of topic modeling studies. We summarized five normative arguments that address critical issues in theory building and testing, research design, and reliability and validity assessments. Using these normative arguments as criteria, we systematically reviewed 105 communication studies that applied topic modeling. We identified gaps and missed opportunities in previous studies and discussed potential pitfalls for the field.

Increasing communication studies have applied topic modeling to analyze texts from digital media. As an unsupervised machine learning method, topic modeling helps researchers identify latent structures within a large volume of texts (DiMaggio et al., 2013). The method not only provides a lens for scholars to identify new concepts, but also serves as an inductive analysis tool to generate potential hypotheses (Grimmer et al., 2021; Margolin, 2019). As the text data largely unexplored before digital media became available to scholars, topic modeling seems to represent an effective and innovative tool for many communication researchers (van Atteveldt & Peng, 2018).

Despite the methodological strength, empirical studies applying topic modeling face a critical question: what are meaningful theoretical contributions? It is perhaps convenient for communication researchers to employ topic modeling to describe themes in a massive amount of text data without having much prior knowledge. A simple inductive research design seems prevalent in topic modeling studies, despite its potential to propose or test a causal relationship. Topic modeling thus gives readers the impression of providing "a mere descriptive exploration of the corpus when employed on its own" (Törnberg & Törnberg, 2016, p. 418). Moreover, procedures for performing topic modeling have not been fully standardized yet (Maier et al., 2018). The lack of reporting methodological details in previous studies makes it difficult to validate the reported findings and adequately replicate the same findings. As many easy-to-use software packages become accessible, topic modeling can be overused without making theory-informed reasoning or close validation.

To understand how communication researchers have utilized topic modeling, we systematically reviewed 105 empirical studies published in major communication journals from 2009 to 2021. We focus on three challenges that often appear in critiques of topic modeling studies: (1) theory building and testing, (2) research design, and (3) assessments of reliability and validity. We first summarize normative arguments from the current literature in topic modeling and computational social science

CONTACT Yingying Chen yychen1911@hotmail.com Renmin University of China, the School of Journalism and Communication, Mingde Building, 59 Zhongguancun Street, Haidian District, Beijing 100872, China

to address critical issues in each challenge. Then, we use the normative arguments as criteria and conduct a quantitative content analysis of the communication studies using topic modeling. While previous reviews of topic modeling studies focused on post hoc validations of extracted topics (e.g. Maier et al., 2018; Ying et al., 2021), our study attempts not only to review such methodological details but also to highlight the importance of theoretical underpinning and research design of topic modeling studies, which together can enhance their theoretical contributions.

All in all, we intend to show that it takes multiple steps for a study to produce meaningful theoretical contributions using topic modeling. It is not the goal of our systematic review to make a judgment about the value of previous studies. Instead, our goal is to identify critical gaps and missed opportunities in previous research, which will help communication researchers take full advantage of topic modeling and make more useful methodological choices. In particular, it seems highly important to continue the dialogue about the norms and expectations of using topic modeling and other computational text analysis methods properly at this relatively early stage of adopting the methodology (Baden et al., 2022; Grimmer et al., 2022; Maier et al., 2018; Margolin, 2019).

Basic introduction and the strength

Topic modeling is a broad term for computer algorithms that automatically identify latent structures from a large volume of text data. As a popular form of topic modeling,[1] probabilistic topic models such as latent Dirichlet allocation (LDA) estimate the structural patterns in text generation processes, the correlation between themes, and their changes over time based on word occurrence (Blei, 2012). As an abundance of texts from digital media became accessible, communication scholars and other social scientists have used topic modeling as an automatic text analysis tool (see Blei (2012) for more technical introductions of probabilistic topic models).

Topic modeling can be a powerful tool in social science in several aspects. First, it helps researchers gain a quick overview of the major contents from a large volume of text data (DiMaggio et al., 2013; Nelson, 2020). As a dimension-reduction technique, topic modeling transforms a large sample of text into a much smaller set of topics. Second, topic modeling provides a new lens for scholars to identify patterns that would otherwise be undetectable with manual coding alone from a massive amount of texts (DiMaggio et al., 2013). The method can be used as an inductive tool to identify categories that have been largely undiscovered before (Nelson, 2020). Lastly, topic modeling helps communication scholars extract certain meanings from the text data (DiMaggio et al., 2013; Grimmer & Stewart, 2013). Since many theoretical concepts in social science are not directly observable, social scientists have relied on topic modeling as a “text-as-data” method to derive measures of unobservable concepts (e.g., radical rhetoric) from written texts (Ying et al., 2021).

To conduct a topic model, researchers need to preprocess text data and transform it into a matrix (e.g., a document-term matrix) as the model input. Then, researchers pick the best number of topics for a topic model. The primary outputs of topic modeling are (1) topics, each linking to a group of words with higher probabilities than others based on their co-occurrence across documents, and (2) topic proportions of each identified topic in each document. Figure 1 provides an example of the topics and topic probabilities. The words and topic probabilities do not directly reveal the meaning of each topic. Researchers need to interpret and label each topic by carefully inspecting words and documents that are most relevant to each topic. As the procedure of conducting a topic model includes several choices that are seemingly subjective (e.g., the number of topics, the interpretation of topics), computational social scientists have developed a set of norms and expectations for how to use topic modeling properly and why. Being well-informed about the norms and expectations will help communication scholars make better decisions in designing and executing a topic modeling study.

[1] Other non-negative matrix factorization (NMF)-based models (Shi et al., 2018) are also considered a topic modeling technique, which researchers can use to classify documents.