CENTRE FOR EAST ASIAN STUDIES + UNIVERSITY OF TURKU + CEAS@UTU.FI + WWW.UTU.FI/CEAS

东亚研究 | 동아시아학 | 東亜研究

CEAS

CEAS INSIGHTS 2/2021

COMPUTATIONAL APPROACHES TO STUDIES ON CONTEMPORARY CHINA: A BRIEF OVERVIEW

Author: Hui Zhao

INTRODUCTION

The field of Chinese studies is increasingly adopting various computational methods (Lazer et al., 2009) to interrogate social issues in China. This represents a convergence and reciprocal learning between different fields with different ways of thinking about and doing social science research in China (e.g., Hao, Zhu, & Zhong, 2015; Jiang & Fu, 2018). The goal of this paper is to provide a brief overview of selected studies with various digital approaches and to demonstrate how these studies productively help us scrutinize Chinese society in new ways. The paper is directed to scholars who are interested in using computational methods in the study of contemporary China, but who do not know much, if anything, about those methods.

The paper does not aim to sketch a full-blown picture of scholarship in Chinese studies with diverse digital and computational methods. Rather, by scrutinizing selected, exemplifying studies, it aims to tease out and consequently call attention to the fact that computational methods, in various forms, contribute to the advancement of Chinese studies. Hence, this paper also invites inputs and engagements from the readers for furthering the involvement of digital and computational methods in Chinese studies.

In the following, this paper presents selected studies that exemplify digital methods from semantic network analysis to newly developed digital methods such as automated text analysis and supervised and unsupervised machine learning. As an initial step to assess knowledge relating to Chinese studies with computational and digital methods, this study reviewed articles published in journals indexed in the Social Sciences Citation Index (SSCI) and the Arts and Humanities Citation Index (A&HCI). The keyword-screening method was applied in the Web of Science databases in October 2021, and the keywords "computational method" or "digital method" and "China," "China studies" or "Chinese studies" were used in all fields. Then the author examined this list of articles and only included those that adopted computational or digital methods (For a discussion on digital or computational methods, see, e.g., Hampton, 2017; Rogers, 2013, 2019). Finally, the articles have been clustered in terms of their disparate methods. If an article involved more than one method, it was multiple coded.

UNSUPERVISED AND SUPERVISED MACHINE LEARNING

The main difference between unsupervised and supervised machine learning is that the latter involves learning on a labeled dataset (usually pre-assigned or hand-labeled for the training data) in order to evaluate the accuracy of the algorithm. In the case of unsupervised machine learning, the algorithm works by extracting patterns on its own.

Unsupervised machine learning

Topic modeling is the most frequently used unsupervised machine learning method in China studies. Topic modeling (e.g., Roberts, Stewart, Tingley, & Airoldi, 2013) is based on the Latent Dirichlet Allocation (LDA) statistical model, aiming to unearth hidden semantic structures in textual data. LDA treats a document as a vector of word counts (text vectorization) and analyses a topic as a multinomial distribution of words (Ramage, Dumais, & Liebling, 2010). The LDA-based topic modeling is "a statistical tool for the automatic discovery and comprehension of latent thematic structure or topics" (Guo, Vargo, Pan, Ding, & Ishwar, 2016). Several recent studies applied topic modeling to understand various topics in China studies. Some apply topic modeling to understand specific hidden topics via social media platforms.

To explore public attitude toward off-site construction (OSC) in China, Wang, Li, and Wu (2019, for details, see pp. 3–4) adopted the Octopus web scraping technology to collect data from the Sina Weibo microblogging site. Combining this with two methods of computer-aided text analysis (text mining), namely topic modeling and sentiment analysis, the study explored Weibo posts with keywords related to off-site construction such as the commonly used terms "OSC" and "industrialized construction". Wang and his colleagues identified thematic topics in the data (topic modeling) through the three steps of 1) text vectorization, 2) topic modeling with LDA, and 3) topic model visualization to reveal and visualize the Chinese public attitude toward off-site construction based on Weibo. Semantic network analysis, which was also adopted in this study, will be introduced later in this paper.

> *Link to the article:* Wang, Y., Li, H., & Wu, Z. (2019). Attitude of the Chinese public toward off-site construction: A text mining study. *Journal of Cleaner Production*, 238, 117926. https://doi.org/10.1016/j.jclepro.2019.117926

Similarly, Zeng, Chan, & Schäfer (2020) utilized topic modeling to study how the Chinese public discuss artificial intelligence in the WeChat messaging and social networking application and also compared the public discussion with narratives from state-owned media People's Daily.

> *Link to the article:* Zeng, J., Chan, C.-h., & Schäfer, M. S. (2020). Contested Chinese dreams of AI? Public discourse about artificial intelligence on WeChat and people's daily online. *Information, Communication & Society*, 1-22. https://doi.org/10.1080/1369118X.2020.1776372

In addition to public discourse, topic modeling could also be used for distinguishing content in the text. In their recent study, Lu & Pan (2021) developed automated scraping algorithms to collect titles of WeChat posts from official city-government propaganda accounts, implementing a Structural Topic Model (Roberts et al., 2014). Topic modeling was used to examine the proportion of propaganda and non-propaganda content in the titles, and then hand-label identified topics.

> *Link to the article*: Lu, Y., & Pan, J. (2021). Capturing clicks: How the Chinese government uses clickbait to compete for visibility. *Political Communication, 38*(1-2), 23-54. https://doi.org/10.1080/10584609.2020.1765914

Jaros & Pan (2018) took named-entity recognition (NER) methods from computer science to analyze a corpus of nearly two million articles from provincial-level Party-run newspapers. As the name of the method reveals, it tracks names of persons and organizations in text. Thereby, it portrays a picture of the individuals and organizations appearing in the news and their relative dominance in media coverage. As described in their design, NER is able to "identify and tabulate virtually all the mentions of different organizations and individuals appearing in a collection of media texts," thereby allowing "to calculate a given political entity's (or set of entities') share of all named-entity occurrences in the media" (Jaros & Pan, 2018, p. 113).

> *Link to the article*: Jaros, K., & Pan, J. (2018). China's Newsmakers: Official Media Coverage and Political Shifts in the Xi Jinping Era. *The China Quarterly, 233*, 111-136. https://doi.org/10.1017/S0305741017001679

Supervised machine learning

Supervised machine learning is a subcategory of machine learning (ML) in which machines are trained using pre-assigned or hand-labeled training data to classify data and predict outcomes. There are two main types of supervised machine learning problems. One is classification that involves predicting a class label; the other is regression that involves predicting a numerical label (Rashidi, Tran, Betts, Howell, & Green, 2019).

After scraping publicly available posts from the web interface of Weibo's search engine with a focus on Chinese profane words, Song et al. (Song, Kwon, Xu, Huang, & Li, 2021, p. 989) conducted emotional classification with supervised machine learning and "automatically classified each [Weibo] post into one of four emotional categories including happiness, fear, anger, and sadness." More specifically, Song et al. employed a so-called convolutional neural network (CNN) algorithm for machine learning process, given the strengths of CNN in performing sentence classification (Zhang & Wallace, 2015). Once the data was cleaned and tokenized (the text split into smaller units or "tokens)", the Word2vec algorithm model was used. That model "allows to capture syntactic and semantic relationships between words, as well as the context of words in a document" (Song et al., 2021, p. 989). Similar to Song, et al. (2021), Göbel (2021) also adopted the Word2vec model to examine a tokenized corpus of social media protest posts from the Wickedonna Dataset.

Links to the articles:
Song, Y., Kwon, K. H., Xu, J., Huang, X., & Li, S. (2021). Curbing profanity online: A network-based diffusion analysis of profane speech on Chinese social media. *New Media & Society*, 23(5), 982-1003. https://doi.org/10.1177/1461444820905068

Göbel, C. (2021). The Political Logic of Protest Repression in China. *Journal of Contemporary China, 30*(128), 169-185. https://doi.org/10.1080/10670564.2020.1790897

Liu & Liu (forthcoming) employed supervised machine learning for emotion recognition and classification of Weibo posts in the case of the Red-Yellow-Blue kindergarten child abuse scandal (the "RYB scandal" for short, Buckley & Kan, 2017) to find out who are the prominent actors leading the diffusion of emotional messages in China's online activism. The design first took 10 percent of the Weibo posts (6,386 posts) as the training dataset for manually labeling emotions in terms of Ekman's (1994; Ekman & Frieden, 1971) six basic emotions in the RYB scandal. Then the remaining 90 percent (57,477 posts) formed the test dataset for automatic emotion recognition and labeling. Next, to automatically classify and label the test dataset, the DL-CKM Natural Language Processing Engine with the Support Vector Machine algorithm was adopted. In total, 63,607 posts were labeled through the automated emotion recognition process.

Link to the article: Liu, N., & Liu, J. (forthcoming). Leading with hearts and minds: Emotion contagion in China's online activism. *Social Movement Studies.*

A combination of supervised and unsupervised machine learning.

Despite the increase in the use of ML and Artificial Intelligence (AI) for the analysis of both textual and multimedia contents, the accuracy of these analyses is greatly dependent on the quality of the training sets for building the machine learning models. Given this consideration, a tendency emerges to adopt unsupervised methods to the dataset for preliminary insight into the quality of the data. Then the selection and labeling of data for creating training sets would give a better result regarding the application of machine learning when compared with the use of either supervised or unsupervised methods alone.

Guo (2019) employed both supervised and unsupervised machine learning methods to analyze 33,875 online news articles that covered the Two Sessions, China's biggest annual political event, from China's official and commercial news websites in March 2017. The study first utilized topic modeling as an unsupervised machine learning approach to automatically identify salient topics in each media category (e.g., official or commercial news websites and national or regional level news websites) of news coverage. The aim of the topic modeling was to uncover latent thematic structure or topics. Next, supervised machine learning was applied to detect main topics reported in each news article. Three student coders manually coded 2,050 news articles as a training set for building Support Vector Machine (SVM) models in supervised machine learning. The acceptable SVM models were then used to detect the topics in the data set.

Link to the article: Guo, L. (2019). Media agenda diversity and intermedia agenda setting in a controlled media environment: A computational analysis of China's online news. *Journalism Studies, 20*(16), 2460-2477. https://www.tandfonline.com/doi/abs/10.1080/1461670X.2019.1601029?journhttps://doi.org/10.1080/1461670X.2019.1601029

SEMANTIC NETWORK ANALYSIS

Semantic network analysis focuses on frequency, co-occurrence, and distances among words and concepts, which allows researchers to explore a text's embedded meaning (Danowski, 1993; Doerfel & Barnett, 1999). There have been different methods to conceptualize network ties, including traditional content analysis, shared perception, and word association in existing semantic network analysis research (Doerfel & Barnett, 1999). The strength of semantic network analysis is that it extends beyond the standard content analysis of texts and frequencies of concepts, reveals the manifest meaning structure of the text and indirectly represents the collective cognitive structure among the text's creators (Danowski, 1993).

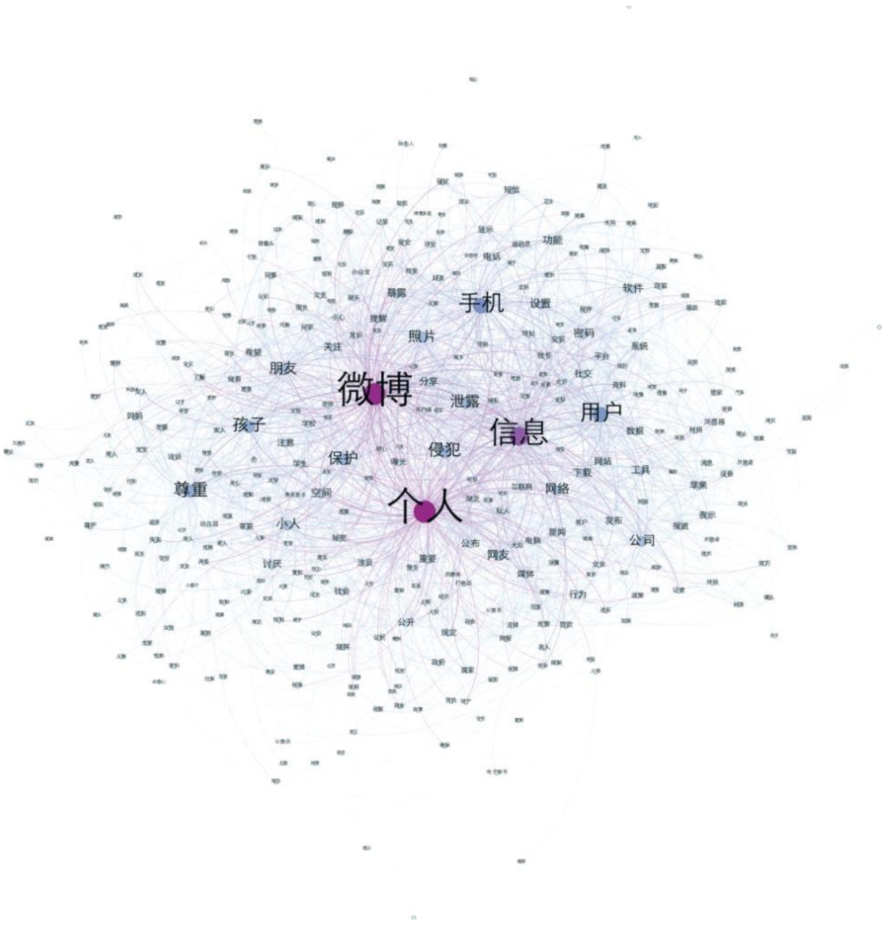

Figure 1. The semantic network of privacy on Sina Weibo (Yuan, Feng & Danowski, 2013)

Yuan, Feng, and Danowski (2013) is one of the first studies that applied semantic network analysis to China-related topics in communication studies. Yuan and her colleagues performed a semantic network analysis of 18,000 posts on Weibo to depict the meanings and boundaries of privacy in contemporary China. By examining nodes linked together with the co-occurrence of different concepts, the semantic network generated in their study identified major privacy-related concepts in the text data to illustrate the important dimensions of privacy discourse on Weibo. Further, the authors utilized conventional discourse analysis to explain the meaning of each cluster with more detailed examples of the postings.

More recently, Liu & Zhao's (2021) study employed the word association (concept co-occurrence) method in semantic network analysis to examine news media frames of Covid-19, technology, and privacy in mainland China. Different from the study of Yuan et al. (2013) on public discourse, the focus of Liu & Zhao was media discourse. They operationalized media frames as complex patterns of associations between different concepts. In this regard, semantic networks derived from the occurrences and co-occurrences of concepts are suitable for investigating media frames. More specifically, they took the word association (concept co-occurrence) method in semantic network analysis to map the relationships among words by indexing pairs of concepts related to state, technology, and privacy.

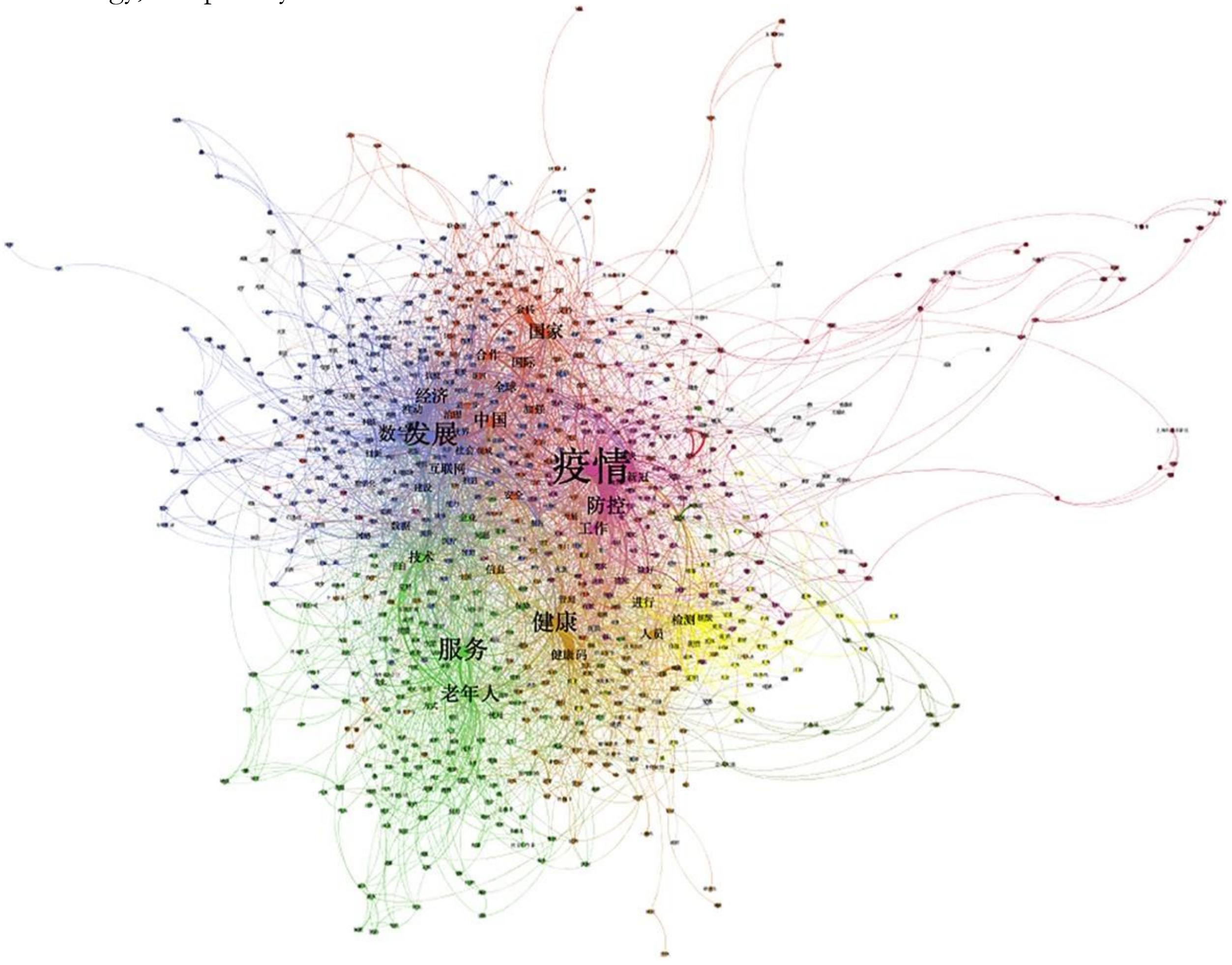

Figure 2. Semantic network of media coverage of Covid-19 (Liu & Zhao, 2021)

Both Yuan et al. (2013) and Liu & Zhao (2021) conducted the analyses using WORDij 3.0 (Danowski, 1993), a computer program that counts the frequency of each word and the co-occurrence of word pairs (http:/wordij.net). In the visualization part of these two studies, Gephi (https://gephi.org/) was also employed to visualize the semantic network and the word cluster.

Links to the articles:

Yuan, E. J., Feng, M., & Danowski, J. A. (2013). "Privacy" in semantic networks on Chinese social media: The case of Sina Weibo. *Journal of Communication, 63*(6), 1011-1031. https://doi.org/10.1111/jcom.12058

Liu, J., & Zhao, H. (2021). Privacy lost: Appropriating surveillance technology in China's fight against COVID-19. *Business Horizons*. *64*(6), 743-756. https://doi.org/10.1016/j.bushor.2021.07.004

ACKNOWLEDGEMENT
The author wishes to thank Outi Luova for the invitation for writing this timely and relevant topic. The author also thanks Jun Liu, Hermann Aubie and Silja Keva for their intellectual input in the development of this paper.

FUNDING
The publication of this paper is funded by the Joel Toivola Foundation.

AUTHOR NOTE
Hui Zhao, PhD, is a senior lecturer in the Department of Strategic Communication at Lund University, Sweden. Her current research includes digital media, social change, crisis communication, and organizational communication. Hui Zhao can be contacted via email: hui.zhao@isk.lu.se.

REFERENCES

Buckley, C., & Kan, K. (2017, Nov. 25). Beijing Kindergarten Is Accused of Abuse, and Internet Erupts in Fury. *New York Times,* p. 10. Retrieved from https://www.nytimes.com/2017/11/24/world/asia/beijing-kindergarten-abuse.html

Danowski, J. A. (1993). Network Analysis of Message Content. In G. Barnett & W. Richards (Eds.), *Progress in Communication Sciences XII* (pp. 197-222). Ablex: Norwood.

Doerfel, M. L., & Barnett, G. A. (1999). A Semantic Network Analysis of the International Communication Association. *Human Communication Research, 25*(4), 589-603.

Göbel, C. (2021). The Political Logic of Protest Repression in China. *Journal of Contemporary China, 30*(128), 169-185.

Guo, L. (2019). Media agenda diversity and intermedia agenda setting in a controlled media environment: A computational analysis of China's online news. *Journalism Studies, 20*(16), 2460-2477.

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism & Mass Communication Quarterly, 93*(2), 332-359.

Hampton, K. N. (2017). Studying the digital: Directions and challenges for digital methods. *Annual Review of Sociology, 43*, 167-188.

Jaros, K., & Pan, J. (2018). China's Newsmakers: Official Media Coverage and Political Shifts in the Xi Jinping Era. *The China Quarterly, 233*, 111-136.

Liu, J., & Zhao, H. (2021). Privacy lost: Appropriating surveillance technology in China's fight against COVID-19. *Business Horizons.*

Liu, N., & Liu, J. (Forthcoming). Leading with hearts and minds: Emotion contagion in China's online activism. *Social Movement Studies.*

Lu, Y., & Pan, J. (2021). Capturing clicks: How the Chinese government uses clickbait to compete for visibility. *Political Communication, 38*(1-2), 23-54.

Ramage, D., Dumais, S., & Liebling, D. (2010). *Characterizing microblogs with topic models.* Paper presented at the Fourth international AAAI conference on weblogs and social media.

Rashidi, H. H., Tran, N. K., Betts, E. V., Howell, L. P., & Green, R. (2019). Artificial intelligence and machine learning in pathology: the present landscape of supervised methods. *Academic pathology, 6*, 2374289519873088.

Roberts, M. E., Stewart, B. M., Tingley, D., & Airoldi, E. M. (2013). *The structural topic model and applied social science.* Paper presented at the Advances in neural information processing systems workshop on topic models: computation, application, and evaluation.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., . . . Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58*(4), 1064-1082.

Rogers, R. (2013). *Digital Methods.* Cambridge: The MIT press.

Rogers, R. (2019). *Doing digital methods*: Sage.

Song, Y., Kwon, K. H., Xu, J., Huang, X., & Li, S. (2021). Curbing profanity online: A network-based diffusion analysis of profane speech on Chinese social media. *New Media & Society, 23*(5), 982-1003.

Wang, Y., Li, H., & Wu, Z. (2019). Attitude of the Chinese public toward off-site construction: A text mining study. *Journal of Cleaner Production, 238*, 117926.

Yuan, E. J., Feng, M., & Danowski, J. A. (2013). "Privacy" in semantic networks on Chinese social media: The case of Sina Weibo. *Journal of Communication, 63*(6), 1011-1031. doi:10.1111/jcom.12058

Zeng, J., Chan, C.-h., & Schäfer, M. S. (2020). Contested Chinese dreams of AI? Public discourse about artificial intelligence on WeChat and people's daily online. *Information, Communication & Society*, 1-22.

Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820.*