Topic modelling and qualitative textual analysis

Karoliina Isoaho, Daria Gritsenko (University of Helsinki)

The pursuit of big data is transforming qualitative textual analysis—a laborious activity that has conventionally been executed manually by researchers. Access to data of unprecedented scale and scope has created a need to both analyse large data sets efficiently and react to their emergence in a near-real-time manner (Mills, 2017). As a result, research practices are also changing. A growing number of scholars have experimented with using machine learning as the main or complementary method for text analysis. Even if the most audacious assumptions 'on the superior forms of intelligence and erudition' of big data analysis are today critically challenged by qualitative and mixed-method researchers (Mills, 2017: 2), it is imperative for scholars using qualitative methods to consider the role of computational techniques in their research (Janasik, Honkela and Bruun, 2009). Social scientists are especially intrigued by the potential of topic modelling (TM), a machine learning method for big data analysis (Blei, 2012), as a tool for analysis of textual data.

This research contributes to a critical discussion in social science methodologies: how topic modeling can concretely be incorporated into existing processes of qualitative textual analysis and interpretation. Some recent studies paid attention to the methodological dimensions of TM vis-à-vis textual analysis. However, these developments remain sporadic, exemplifying a need for a systematic account of the conditions under which TM can be useful for social scientists engaged in textual analysis. This paper builds upon the existing discussions, and takes a step further by comparing the assumptions, analytical procedures and conventional usage of qualitative textual analysis methods and TM. Our findings show that for content and classification methods, embedding TM into research design can partially and, arguably, in some cases fully automate the analysis. Discourse and representation methods can be augmented with TM in sequential mixed-method research design.

We outline avenues for TM both in embedded and sequential mixed-method research design. This is in line with previous work on mixed-method research that has challenged the traditional assumption of there being a clear division between qualitative and quantitative methods. Scholarly capacity to craft a robust research design depends on researchers' familiarity with specific techniques, their epistemological assumptions, and good knowledge of the phenomena that are being investigated to facilitate the substantial interpretation of the results. We expect this research to help identify and address the critical points, thereby assisting researchers in the development of novel mixed-method designs that unlock the potential of TM in qualitative textual analysis without compromising methodological robustness.

Blei, D. M. (2012) ‘Probabilistic topic models’, Communications of the ACM, 55(4), p. 77.

Janasik, N.,Honkela, T. and Bruun, H. (2009) ‘Text Mining in Qualitative Research’, Organizational Research Methods, 12(3), pp. 436–460.

Mills, K. A. (2017) ‘What are the threats and potentials of big data for qualitative research?’, Qualitative Research, p. 146879411774346.