ANALYTIC ESSAY

Methods Content Analysis: A Role in Applied Health Research

Peter Cahill[1]
1. Rehabilitation Science, McMaster University, cahillp@mcmaster.ca

ABSTRACT

Text data is highly information-rich and accessing this information would greatly benefit applied health researchers and decision makers. Text data can be viewed as both qualitative and quantitative by the researcher. When both the quality and quantity of the data can be informative, a rigorous mixed methods approach is necessary to make best use of available analysis techniques to yield high quality inferences. In this analytic essay, a sketch of a suggested mixed methods content analysis method is provided, combining the rich interpretive power of close readings of text data by researchers with robust quantitative modelling via machine learning. This mixed methods content analysis method appears promising for public health systems.

Received: 16/03/2022 **Accepted:** 12/08/2022 **Published:** 01/12/2022

Keywords: Content analysis; machine learning; topic modelling; mixed methods; text analysis

INTRODUCTION

Humans use a plethora of communication methods to exchange messages, such as through language, facial expressions, and gestures. Among such methods, language likely carries the majority of information exchanged. In many societies, this linguistic information is frequently represented by text data, encoded in one of the many writing systems currently used around the globe. Here, text data refers to any written encoding of linguistic information. While the quantity of text data generated by human activity has exploded, digital and networking technologies have made it feasible for researchers to access and collate this data (Blei, 2012; Albalawi et al., 2020). Consequently, methods and technologies aimed to analyze and interpret these vast quantities of rich data are timely and promising (Grimmer & Stewart, 2013). Language is too rich to be fully modelled and, thus, the analysis of text data is highly complex; nevertheless, it offers enormous potential to unlock vast amounts of additional information (Grimmer & Stewart, 2013; Blei, 2012).

Health systems produce vast amounts of textual data in their daily operations—for example, media releases, strategic planning documents, and patient feedback forms. Further, individuals themselves may discuss their own health in many text-based forums, such as social media platforms, providing data regarding what people say about their own health and their understanding of public health and healthcare services. These various sources are likely to contain useful information for applied health researchers to draw inferences about the health of certain communities and health systems.

CONTENT ANALYSIS APPROACHES

Analysis of textual data, particularly via qualitative methods, has a rich tradition within health research. For example, qualitative content analysis has become particularly important and widespread within nursing and public health literatures (Elo & Kyngäs, 2008). Content analysis allows the researcher to describe and quantify the content contained therein, as well as sort and categorize

the data into a smaller number of meaningful units with shared meaning (Elo & Kyngäs, 2008). In this technique, the researcher carefully examines the text data, usually applying codes in iterative rounds. The method is grounded in a close reading of the data, using rich human interpretive abilities to identify commonalities among sections of text data, to then reorganize and reinterpret this data along new organizing qualities perceived by the researcher (Eickhoff & Wieneke, 2018).

Content analysis may involve inductive coding (i.e., conventional content analysis), deductive coding using established frameworks or theories (i.e., directed content analysis), as well as a combination of qualitative and quantitative approaches to the language used in the text (i.e., summative content analysis; Hsieh & Shannon, 2005). These techniques range from more numeric approaches (counting word frequency), to assessing the manifest content of the text (what does the text literally and explicitly say), to more interpretative approaches (searching for communicative intent and interpretation) where the researcher investigates the latent meaning of the text (Lindgren et al., 2020).

Content analysis, in comparison to other common qualitative data analysis techniques, is partially unique in that it allows the researcher to cross between qualitative and quantitative readings of the data (Vaismoradi et al., 2013). Many other qualitative data analysis methods would not be paradigmatically consistent with quantifying text data (Isoaho et al., 2021). The ability of content analysis to accommodate the quantification of text data has the subsequent methodological benefit of allowing a close reading of such data by researchers to be supplemented with recent advances in machine learning modelling (Isoaho et al., 2021). These machine learning techniques can offer a "distant reading (Eickhoff & Wieneke, 2018, p. 906)" of the data that can rapidly reduce vast amounts of data to a smaller number of underlying core components (Grimmer & Stewart, 2013). Many researchers have concluded that, using these techniques, machine learning algorithms can indeed read the text and produce a similar analysis to a human reader, which may have implications on the efficiency and use of societal resources when analyzing the content of text data (Isoaho et al., 2021; Lucas et al., 2015).

TOPIC MODELLING

Topic models are a family of machine learning techniques that help identify the main themes or topics present within the content of large sets of text data (Blei, 2012). They are unsupervised machine learning techniques, meaning that they perform exploratory or inductive analyses of the data (Grimmer & Stewart, 2013). Although this is only one subtype of potential machine learning applications to text data (see Grimmer & Stewart, 2013 for a discussion and taxonomy of methods), topic modelling methods have found enormous popularity and potential in applied sciences of human behaviour, particularly policy studies and political science (Isoaho et al., 2021; Lucas et al., 2015; Roberts et al., 2014, 2016). Two popular topic modelling techniques are outlined below—specifically, Latent Dirichlet Allocation (LDA; Blei, 2012) and Structured Topic Modelling (STM; Roberts et al., 2016).

LDA is a popular topic modelling algorithm, which may be related to its performance with short text documents, such as social media posts. In LDA, documents are assumed to be generated by a stochastic process, with each document composed of one or more topics, and each topic of certain words (Albalawi et al., 2020; Blei, 2012). To put it differently, a document is formed by randomly drawing topics, and a topic is formed by randomly drawing words (Blei, 2012; Grimmer & Stewart, 2013). Although this data generation mechanism is clearly *not* an accurate model, as we do not compose documents or texts by randomly generating and combining words, these model assumptions appear to perform well at retrieving the semantic content (i.e., the meaning of words) of text data (Grimmer & Stewart, 2013).

STM is a similar topic modelling algorithm, differing from LDA primarily in its incorporation of document metadata (Roberts et al., 2016). This approach has been used to explore the effect of document attributes on topic distributions (Roberts et al., 2016). For example, it has been used to compare and contrast online political discourse surrounding actions taken by the United States Federal Government in Arabic- and Chinese-speaking communities (Lucas et al., 2015). STM has also been used to explore how political alignments are associated with different topics (Roberts et al., 2014). For both approaches, a close reading of the text by human readers is

considered the gold standard to validate that topics retrieved by the algorithm are meaningful and justifiable (Grimmer & Stewart, 2013; Roberts et al., 2016).

Topic modelling approaches are efficient and effective at extracting high quality inferences regarding the content of the text data relevant to public health. For example, researchers may be interested in discovering the topics discussed on a social media platform around a specific condition and therefore apply LDA. Another suggestion would be to collect text data from different stakeholder groups on a recently implemented public health intervention through interviews, open-ended surveys, or focus groups. Using STM would allow an exploration not only of the topics present within the data, but also would facilitate the inclusion of relevant demographic characteristics to estimate differences among stakeholders. To illustrate, clinicians may devote substantially more time to work process issues, whereas patients and the general public may focus on accessibility and perceived overall efficacy.

Although these machine learning techniques are promising, the ability to perform a topic model does not necessarily justify its use. At this point, the relevant question is how to best make sense of the myriad of techniques available for analyzing the content of text data and to select the most appropriate method for a research question that would yield high-quality inferences.

MIXED METHODS CONTENT ANALYSIS

The selection of a content analysis method will clearly depend on the research question, as well as the assumptions that the researcher makes about the text data. Text data can be treated as qualitative, quantitative, or both (Eickhoff & Wieneke, 2018). If the researcher assumes that the text information represents *quality* accessible via human interpretation, then a strictly qualitative content analysis is appropriate, and any quantification (such as counting code frequency) seems incongruent. Alternatively, if the researcher approaches the text as representing *quantity*, a quantitative content analysis may be more suitable, as inferences about the meaning of the text may be limited. While there may be merit to both of these perspectives, in applied health research, it is likely that most projects working with text data will be drawn to both the quality and quantity of the data. For example, applied health researchers may be interested in both the meaning of the text, as well as whether certain topics only appear in text from specific stakeholder groups, such as marginalized populations. Removing either quality or quantity from the research would be a lost opportunity for applied health research in many cases. However, careful consideration of how to combine the analysis of these aspects of the data is necessary. Indeed, a mixed methods content analysis may be required.

Summative content analysis—where the researcher attends both to the frequency of terms within the text, as well as their meaning—already crosses into mixed methods, as researchers using this approach consider the data as both qualitative and quantitative (Hsieh & Shannon, 2005). Therefore, it should be possible to develop a robust mixed method content analysis using the paradigmatic assumptions of mixed methods to bolster inference quality. Improving inference quality is analogous to *trustworthiness* in qualitative research and *validity* in quantitative approaches (Onwuegbuzie et al., 2011).

Used in careful combination, a summative content analysis articulated as a mixed method may be valuable to improve applied health research working with text data, using the complementary strengths of qualitative and quantitative approaches to build the analysis (Hsieh & Shannon, 2005). Consistent with qualitative content analysis recommendations, the researcher should become familiarized with the content of the data as a first step (Elo & Kyngäs, 2008). Having a qualitative sense of the content in the data may greatly facilitate the interpretation of quantitative model results. Next, the researcher can then *quantitize* the data by fitting an appropriate topic model (Teddlie & Tashakkori, 2009). As human interpretation suffers from many cognitive biases affecting quantification tasks, such as the base rate fallacy and availability heuristics, the use of topic models has the benefit of removing brute force counting of terms and their co-occurrence by human coders (Berthet, 2021; Blumenthal-Barby & Krieger, 2015; Kliegr et al., 2021; Saposnik et al., 2016). Instead, this task is completed in a reproducible and auditable manner by the machine. However, topic model literature recommends human interpretation of the final model as a gold standard for evaluating the model (Grimmer & Stewart, 2013; Roberts et al., 2016). Consequently, the researcher should *re-qualitize* the data (Teddlie & Tashakkori, 2009). Specifically, the analyst should complete a close reading of at least some of the raw data, using the

topics generated by the quantitative model to tag relevant excerpts, and qualitatively judge the topic model solution, returning to fit another model if necessary.

CONCLUSION

A mixed methods summative content analysis, with topic modelling embedded between rounds of qualitative analysis is an iterative sequential mixed method design which would make best use of the strengths of qualitative interpretation and quantitative analysis of text data (Teddlie & Tashakkori, 2009). Such an approach is likely to yield meaningful inferences for research in public health and health systems, allowing researchers and decision makers access to multiple aspects of the incredibly rich information contained within text data. Consequently, mixed methods content analysis may have an important role to play within applied health research.

Additionally, there may be further mixed methods approaches which combine other types of text analysis with various machine learning algorithms to extract high-quality, impactful inferences from large quantities of text data. This is a promising area of methodological and technological innovation, allowing applied health researchers to unlock powerful insights from text data. Here, I have sketched out only one encouraging approach to unlock these insights from text quality and quantity.

REFERENCES

Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence, 3*(00042). https://doi.org/10.3389/frai.2020.00042

Berthet, V. (2021). The measurement of individual differences in cognitive biases: A review and improvement. *Frontiers in Psychology, 12*(630177). https://doi.org/10.3389/fpsyg.2021.630177

Blei, D. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84. https://doi.org/10.1109/MSP.2010.938079

Blumenthal-Barby, J. S., & Krieger, H. (2015). Cognitive biases and heuristics in medical decision making: A critical review using a systematic search strategy. *Medical Decision Making, 35*(4), 539–557. https://doi.org/10.1177/0272989X14547740

Eickhoff, M., & Wieneke, R. (2018). Understanding topic models in context: A mixed-methods approach to the meaningful analysis of large document collections. *Proceedings of the 51st Annual Hawaii International Conference on System Sciences*, 903–912. https://doi.org/10.24251/hicss.2018.113

Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing, 62*(1), 107–115. https://doi.org/10.1111/j.1365-2648.2007.04569.x

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research, 15*(9), 1277–1288. https://doi.org/10.1177/1049732305276687

Isoaho, K., Gritsenko, D., & Mäkelä, E. (2021). Topic modeling and text analysis for qualitative policy research. *Policy Studies Journal, 49*(1), 300–324. https://doi.org/10.1111/psj.12343

Kliegr, T., Bahník, Š., & Fürnkranz, J. (2021). A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence, 295*(103458). https://doi.org/10.1016/j.artint.2021.103458

Lindgren, B. M., Lundman, B., & Graneheim, U. H. (2020). Abstraction and interpretation during the qualitative content analysis process. *International Journal of Nursing Studies, 108*. https://doi.org/10.1016/j.ijnurstu.2020.103632

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis, 23*(2), 254–277. https://doi.org/10.1093/pan/mpu019

Onwuegbuzie, A. J., Johnson, R. B., & Collins, K. M. T. (2011). Assessing legitimation in mixed research: A new framework. *Quality and Quantity, 45*(6), 1253–1271. https://doi.org/10.1007/s11135-009-9289-9

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58*, 1064–1082) https://doi.org/10.1111/ajps.12103

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association, 111*(515), 988–1003.

Saposnik, G., Redelmeier, D., Ruff, C. C., & Tobler, P. N. (2016). Cognitive biases associated with medical decisions: A systematic review. *BMC Medical Informatics and Decision Making, 16*(1), 1–14. https://doi.org/10.1186/s12911-016-0377-1

Teddlie, C., & Tashakkori, A. (2009). Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioural sciences. SAGE Publications, Inc.

Vaismoradi, M., Turunen, H., & Bondas, T. (2013). Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. *Nursing and Health Sciences, 15*(3), 398–405. https://doi.org/10.1111/nhs.12048