Anne Kroon,
Faculty of Social and Behavioural Sciences

UNIVERSITEIT VAN AMSTERDAM

Take Home Exam:

Big Data and Automated Content Analysis

Helge Moes
11348801
11348801@uva.nl

Date: 11.05.2023

1 Short Essay: Comment on Methodological Choices

Computational techniques have transformed the social sciences in recent decades, empowering researchers to analyze complex data sets, create sophisticated models, and gain novel insights into human behavior and social phenomena. Kitchin (2014), van Atteveldt and Peng (2018) highlight this matter in their literature. In addition, Kitchin (2014) states that the development of Big Data and new data analytics presents a possibility of reframing the epistemology of science, social science, and humanities by enabling new approaches to data generation and analyses.

This essay explores the impact of computational techniques on the social sciences, from the opportunities and challenges presented by these changes to how social sciences have benefitted from these technological developments. Continued innovation in computational techniques holds the potential to revolutionize the study and understanding of the social world. This is examined based on an empirical study by Möller, Kühne, Baumgartner and Peter (2019): *Exploring User Responses to Entertainment and Political Videos: An Automated Content Analysis of YouTube.*

This study analyzed social information about entertainment and political videos on YouTube. The study aimed to examine the role of social identity and intergroup contact in online discussions surrounding entertainment and political videos on YouTube. An automated content analysis was applied as a computational technique in order to collect (dis)likes, views, and 39,602 comments of videos that were selected from the 50 YouTube channels with the highest total number of viewers (Möller et al., 2019). A total of 649 videos that were most watched on each channel were analyzed, with 148 categorized as political entertainment videos and 544 as entertainment videos (Möller et al., 2019).

Additionally, the data on the videos were collected using a Python 3.5 script that employed the Google API (Möller et al., 2019). Information was collected on the videos' title, description, number of likes and dislikes, number of views, and number of comments. The first five pages of comments posted in response to each video were also collected. The information of 101,889 comments written in response to 692 videos was collected (Möller et al., 2019). The analysis focused on the content of the comments posted in response to the videos, with a particular focus on the use of group identity and the extent of intergroup contact evident in the comments.

The amount of information that was collected for this research was challenging to gather through manual means. Therefore, the implications of the Automated content analysis method is that it is faster and more objective than manual content analysis as it relies on algorithms and statistical models to identify patterns in data (Kitchin,2014). However, the use of big data also presents challenges such as ensuring validity, reliability, and ethical considerations (van Atteveldt & Peng, 2018). It can handle larger volumes of data but may miss important nuances that only a human eye can detect and may not be able to handle non-textual data.

On the other hand, manual content analysis is often more subjective and time-consuming than automated content analysis, but it allows for greater flexibility in interpreting data and identifying patterns that may not be detected by automated methods (van Atteveldt & Peng, 2018). The choice between manual and automated content analysis depends on the research question, the type of data being analyzed, and the resources available (van Atteveldt & Peng, 2018). Yet, manual content analysis may be more appropriate for a deep understanding of data and complex coding schemes, while automated content analysis may be more appropriate for analyzing large volumes of data quickly.

A traditional approach Möller et al. (2019) could have made, was to conduct the content analysis entirely manual. In the study by Möller et al. (2019), the videos were hand-coded as either entertainment or political, with entertainment videos being defined as videos that can offer entertainment either while watching the videos or as a direct consequence of watching them, and political videos being defined as videos that discuss topics which are relevant for, and have a direct or indirect influence on, a considerable number of individuals within a society and therefore create a need for discussion or action. Although this is challenging to execute by automated means solely, such as by Socialblade, human intervention proves to be invaluable (Möller et al., 2019).

Therefore, a traditional approach of manually conducting a content analysis can always enrich the research. Nevertheless, this results in a lot of time and resources to analyze information of 101,899 comments. Consequently, a combination of these two methods enables the research to have reliable data. This shows that Big Data presents both opportunities and challenges, including a skills deficit for analyzing and making sense of such data, and the creation of an epistemological approach that enables post-positivist forms of computational social science (Kitchin, 2014).

3 Plan an analysis

Research Question:

To what extent can we identify the factors that contribute to the success of hotels in terms of customer satisfaction on Tripadvisor?

Background:

A hotel's success highly depends on the customer's satisfaction. Therefore, understanding the factors that affect customer satisfaction is fundamental for hotels to improve their services and attract more customers. With the availability of large datasets by Tripadvisor containing information about different hotels and their attributes, it is possible to analyze these factors.

Approach:

To answer the research question, a large dataset will be used containing information of potentially tens of thousands of hotels that are associated with Tripadvisor. Moreover, additional variables of the hotels are requested to identify the factors that influence customer satisfaction. The dataset shall include variables such as hotel location, pricing, star rating, amenities, customer reviews, ratings by customers, etc. Moreover, these quantitative variables can be examined through descriptives and regression analyses in order to predict the customer satisfaction score for each hotel (Hossen et al., 2021).

For analyzing this qualitative information, Natural Language Processing is considered. Natural Language Processing (NLP) entails the application of computational techniques to analyze, understand, and derive meaning from human language (Hossen et al., 2021). Among the most widely used NLP techniques is sentiment analysis, which aims to identify the opinions or sentiments expressed in a given in the reviews. This shall grant insights of the customers that indicates the level of satisfaction (Hosson et al., 2021). Finally, the findings of this research shall provide improvements of the hotels and their services to attract more customers.

Feasibility:

This study can be conducted with limited resources. The dataset can be obtained from public available sources or it can be requested from Tripadvisor. The analysis can be conducted using open-source statistical software such as Python or R.

Creativity:

This study can be considered to be creative in that it combines various methods such as descriptives, regression analyses and sentiment analysis to identify the most important factors that contribute to customer satisfaction in hotels. It also provides insights and recommendations for hotels to improve their services and attract more customers, which can be valuable for the industry. By conducting this study, it can improve our understanding of the hotel industry and contribute to its success.

References

Hossen, Md. S., Jony, A. H., Tabassum, T., Islam, Md. T., Rahman, M. M., & Khatun, T. (2021). Hotel review analysis for the prediction of business using deep learning approach. *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, 1489–1494. https://doi.org/10.1109/ICAIS50930.2021.9395757

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 2053951714528481. https://doi.org/10.1177/2053951714528481

Möller, A. M., Kühne, R., Baumgartner, S. E., & Peter, J. (2019). Exploring User Responses to Entertainment and Political Videos: An Automated Content Analysis of YouTube. *Social Science Computer Review*, *37*(4), 510–528. https://doi.org/10.1177/0894439318779336

van Atteveldt, W., & Peng, T.-Q. (2018). When Communication Meets Computation: Opportunities, Challenges, and Pitfalls in Computational Communication Science. *Communication Methods and Measures*, *12*(2–3), 81–92. https://doi.org/10.1080/19312458.2018.1458084