Selected Papers of #AoIR2023:
The 24th Annual Conference of the
Association of Internet Researchers
Philadelphia, PA, USA / 18-21 Oct 2023

WITH OR WITHOUT THE CROWD? THE INFLUENCE OF CODER CHARACTERISTICS ON CODING DECISIONS COMPARING CROWDWORKERS AND TRADITIONAL CODERS.

Julia Niemann-Lenz
University of Hamburg

Anja Dittrich
Hanover University of Music, Drama, & Media

Jule Scheper
Hanover University of Music, Drama, & Media

Research Objective

Standardized content analysis is a prominent method in internet research; it is used, for example, to evaluate topics, tendencies, and sentiment of textual digital trace data, such as social media posts or comments. The importance of the method has even increased due to the demand of high-quality training data for text mining approaches based upon supervised machine learning. However, standardized content analyses are very time- and resource-intensive, mainly due to the high demand for human resources. Here, the process of coding[1] is particularly critical: To obtain sufficient reliability, it is essential that all coders use the codebook in the same way (Krippendorff, 2002). However, coders code under the conditions of a specific context, which is shaped by their individual characteristics, the coding situation, and the material to be coded (Degen, 2015). Notwithstanding the importance of the coding process, little research has been done on how coding should be optimally designed and how personal characteristics of the coders influence the coding results.

[1] We use the term "coding" because it is common terminology in literature (e.g. Krippendorff 2002). However, the considerations and results provided by our study are applicable to "annotating" as well, which is a corresponding task in computer science.

Suggested Citation (APA): Niemann-Lenz, J., Dittrich, A., Scheper, J. (2023, October). With Or Without The Crowd? The Influence Of Coder Characteristics On Coding Decisions Comparing Crowdworkers And Traditional Coders. Paper presented at AoIR2023: The 24th Annual Conference of the Association of Internet Researchers. Philadelphia, PA, USA: AoIR. Retrieved from http://spir.aoir.org.

The issue raised above is intensified by the fact that an increasing number of studies rely on crowdworkers as coders (e.g., Boxmann-Shabtai 2021; Budak, Goel & Rao 2016; Hornik et al. 2022). This is reasonable, for crowdworkers are cost-effective and flexible in employment. However, in content analysis they bear substantial disadvantages: In addition to ethical aspects—such as the question of appropriate compensation for research work—the use of crowdworkers also exacerbate blind spots in the research process. In particular, the selection of suitable coders and the poor control of the coding process are to be mentioned. The employment of crowdworkers as coders should therefore not be undertaken without reflection. A first study evaluating the use of crowdworkers concludes that they offer a high potential for content analysis and that they do not perform significantly worse than classic offline coders (Lind et al. 2017). However, the reliability coefficients reported there are by no means satisfactory according to common criteria.

Against this background, we investigate comparatively, in a sample of crowdworkers and student coders, which influence coder characteristics have on the validity and reliability of the coding results. Specifically, we ask (a) how the quality of coding with crowdworkers compares to offline coders, and (b) how coder characteristics affect the quality of the data. Our hypotheses relate to the difference between traditional coders and crowdworkers and to the influence of socio-demographic characteristics (age, gender, and formal education), personality traits (need for cognition and emotional sensitivity), and (3) coding tasks of varying difficulty. The results of the study help to eliminate the influence of confounding factors, e.g., by weighting or selecting coders to minimize systematic errors in the future.

Method

To answer our research questions, we employed a mixed method design combining content analysis and survey. For the content analysis we picked a task that is common in internet research: the analysis of tweets. The example topic focused on the current debate about the legalization of abortion in Germany. In addition to its typicality, relevance and topicality, this topic has further merits: It is controversial, so it allows for differing opinions, and it is complex so that variables with varying degrees of complexity can be elicited. Moreover, the analysis of the rather short posts is a task that crowdworkers typically do as part of their job.

The content analysis sample consists of 300 tweets that contain the term "abortion" and actually relate to the topic. The (very short) codebook includes variables that differ in terms of the interpretive effort to be made: (1) positioning (pro-abortion vs. anti-abortion or neutral), (2) emotional valence, (3) references to the current legal situation, and finally, as more formal aspects of the texts, the number of (4) question marks and (5) exclamation marks. The codebook was pretested and optimized in terms of wording. It was intentionally designed to be challenging in order to generate variance in the coding. Initially, all tweets were coded by two researchers who discussed and reconciled their coding decisions. This reference data served as the gold standard.

Data was collected using an online questionnaire as an input mask. The survey contained questions on coder characteristics, short briefings, the codebook, and the

tweets. The comparative study design involved conducting the same content analysis twice: In the first condition, four students who received a training coded the whole sample of 300 tweets. In the second condition, 150 crowdworkers were recruited via the platform Clickworker. They coded 30 tweets each without a training. Depending on their coding, the all coders received feedback every fifth tweet to ensure higher quality of coding decisions. In total, the study includes 28,500 coding decisions. All coders ere adequatly compensated, either in form of credit points (student coders) or with a monetary incentive (crowdworker).

Results

Our data analysis includes systematic comparison of established reliability coefficients (inter coder agreement and Krippendorff's alpha) to obtain a comprehensive picture of the quality of coding in the different conditions. In addition, we matched the coding with the gold standard. Further, we tested the influence of the coders' characteristics on the number of correct coding decisions by regression analysis. Notably, the data comprises a complex dependency structure. The codings are nested in both: coders as well as tweets. Hence, multi-level analysis was applied.

Descriptive findings imply that student coders and crowdworkers differ regarding validity and reliability. Student coders performed better when compared to the gold standard, especially for the pro-abortion vs. anti-abortion categorization (83 vs. 74 percent compliance). Overall, however, coding was found to be difficult regarding validity, as agreement with the gold standard is rather low in both groups (average 62 percent). In contrast, reliability is generally higher (average inter coder agreement: .78; average Krippendoff's alpha = .73). Notably, for the inter coder agreement and even more for Krippendorff's alpha, student coders outperform crowdworkers (Krippendoff's $alpha_{students}$ = .83 vs. Krippendoff's $alpha_{crowdworker}$ = .64). In conclusion, the results advocate the use of traditional coders for greater control of the coding process as well as intensive coding training.

Moreover, multivariate regression analysis indicate that the influence of the categories or variable types of varying difficulty is much stronger than the influence of the coder characteristics. This applies both to agreement with the gold standard and to inter-coder agreement. This result highlights the importance of conducting simple coding tasks, formulating categories precisely and choosing concise operationalizations—especially when working with crowdworkers. Nevertheless, certain coder characteristics also positively determine coding results, this should be considered in the recruitment of coders. Individuals with a high level of formal education and emotional sensitivity are to be preferred. Further implications and recommendations were discussed at the conference.

References

Boxman-Shabtai, L. (2021). Encoding polysemy in the news. *Journalism*, 14648849211045964. https://doi.org/10.1177/14648849211045963

Budak, C., Goel, S., & Rao, J. M. (2016). Fair and balanced? Quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly, 80*(1), 250-271. https://doi.org/10.1093/poq/nfw007

Degen, M. (2015). Codierer-Effekte in Inhaltsanalysen – ein vernachlässigtes Forschungsfeld [Coder effects in content analyses---a neglected field of research]. In W. Wirth, K. Sommer, M. Wettstein & J. Matthes (Eds.), *Qualitätskriterien in der Inhaltsanalyse [Quality criteria in content analysis]* (pp. 78-95). Herbert von Halem.

Hornik, R., Binns, S., Emery, S., Epstein, V. M., Jeong, M., Kim, K., Kim, Y., Kranzler, E. C., Jesch, E., Lee, S. J., Levin, A. V., Liu, J., O'Donnell, M. B., Siegel, L., Tran, H., Williams, S., Yang, Q., & Gibson, L. A. (2022). The Effects of Tobacco Coverage in the Public Communication Environment on Young People's Decisions to Smoke Combustible Cigarettes. *Journal of Communication, 72*(2), 187–213. https://doi.org/10.1093/joc/jqab052

Krippendorff, K. (2002). *Content Analysis. An Introduction to its Methodology* (2nd. Ed.). Sage.

Lind, F., Gruber, M., & Boomgaarden, H. G. (2017). Content Analysis by the Crowd: Assessing the Usability of Crowdsourcing for Coding Latent Constructs. *Communication Methods and Measures, 11*(3), 191–209. https://doi.org/10.1080/19312458.2017.1317338