

Qualitative Coding in the Computational Era: A Hybrid Approach to Improve Reliability and Reduce Effort for Coding Ethnographic Interviews

Zhuofan Li¹ , Daniel Dohan², and Corey M. Abramson¹

Abstract

Sociologists have argued that there is value in incorporating computational tools into qualitative research, including using machine learning to code qualitative data. Yet standard computational approaches do not neatly align with traditional qualitative practices. The authors introduce a hybrid human-machine learning approach (HHMLA) that combines a contemporary iterative approach to qualitative coding with advanced word embedding models that allow contextual interpretation beyond what can be reliably accomplished with conventional computational approaches. The results, drawn from an analysis of 87 human-coded ethnographic interview transcripts, demonstrate that HHMLA can code data sets at a fraction of the effort of human-only strategies, saving hundreds of hours labor in even modestly sized qualitative studies, while improving coding reliability. The authors conclude that HHMLA may provide a promising model for coding data sets where human-only coding would be logistically prohibitive but conventional computational approaches would be inadequate given qualitative foci.

Keywords

computational social science, qualitative methods, machine learning, natural language processing, coding reliability, computational ethnography

Sociologists have argued that there is value in incorporating computational tools into qualitative research, including using machine learning to code qualitative data (Abramson et al. 2018; DiMaggio 2015; Nelson et al. 2021). Yet standard computational approaches do not neatly align with traditional qualitative practices. Researchers have applied machine coding to found data such as written documents. However, qualitative data produced through researcher interactions with human subjects, involving smaller volumes of text, and coded interpretively at a more granular level, such as field notes and interview transcripts, remain common primary data sources in qualitative research and often present barriers to transposing computational social science techniques. We address challenges and illustrate opportunities by integrating BERT, a transfer learning model for sentence-level language understanding (Devlin et al. 2019), into the iterative coding of ethnographic interviews. Our results demonstrate how a hybrid human-machine learning approach (HHMLA) can code interviews at a fraction of the effort of

human-only coding while improving reliability compared with human and machine coding.

Our data consist of 9,673 paragraphs of respondent speech from 87 in-depth interviews collected for a comparative ethnographic study of experiences of advanced cancer. We assess the performance of three coding strategies using the code applied to any paragraph in which “the patient shares any medical history or health information.” Identifying these paragraphs requires contextual interpretation beyond the capability of a dictionary-based search. All data had been previously human-coded using a contemporary iterative

¹University of Arizona, Tucson, AZ, USA

²University of California, San Francisco, San Francisco, CA, USA

Corresponding Author:

Corey M. Abramson, University of Arizona, School of Sociology, 1145 E. South Campus Drive, Social Sciences Building, Room 400, Tucson, AZ 85721, USA.

Email: coreyabramson@arizona.edu



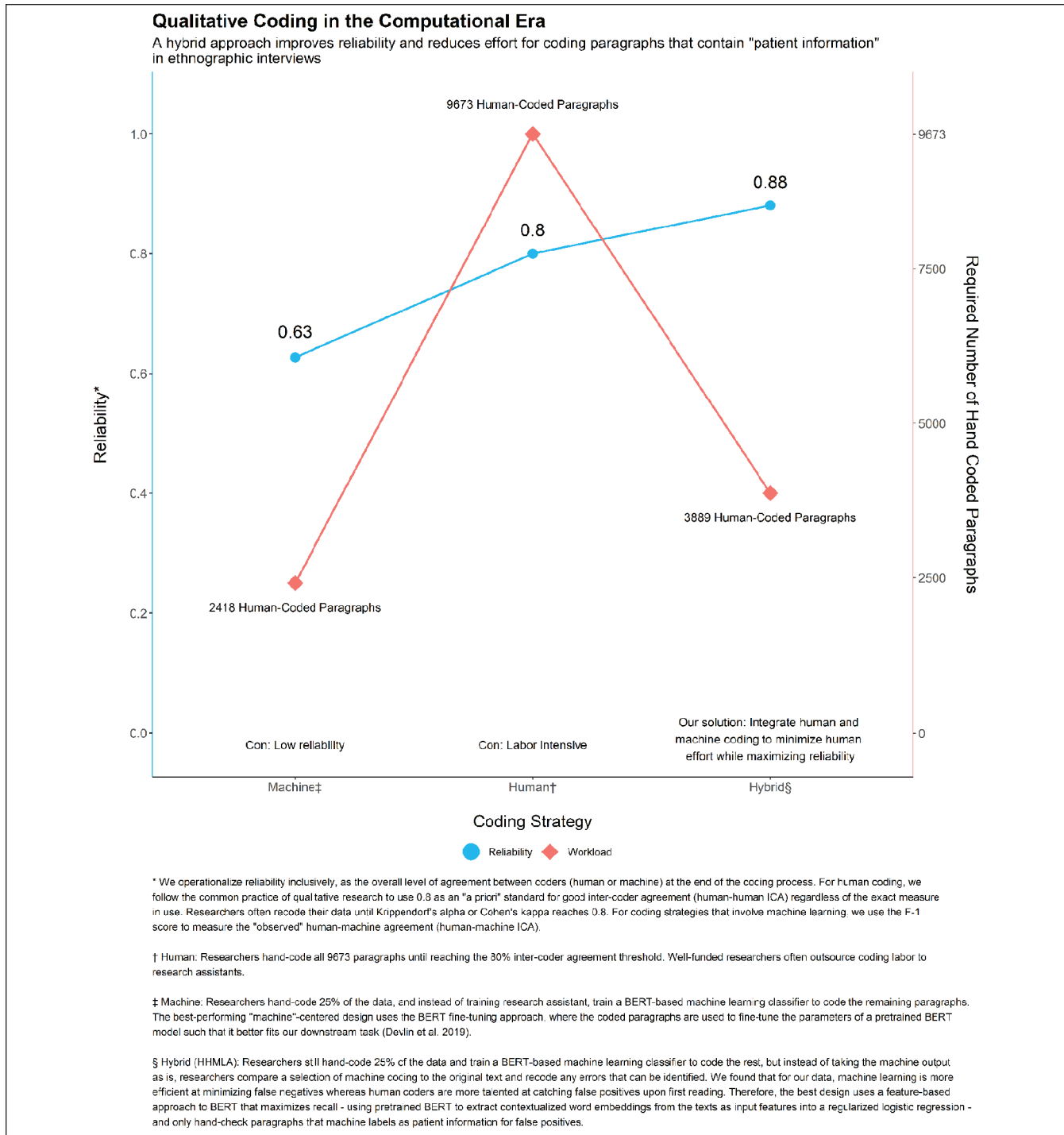


Figure 1. A hybrid approach improves reliability and reduces human effort for coding ethnographic interviews.

approach grounded in analyst interpretation (Deterding and Waters 2021). This provides the baseline “human”-only strategy we use for comparison. For the “machine”-centered strategy, we fine-tuned a pretrained BERT model on 25 percent of the human-coded paragraphs as training data to code the reminder as a series of binary classification tasks. For the

“hybrid” strategy, we combined logistic regression with BERT embeddings, trained this more sensitive algorithm to cast a wider net on potentially relevant paragraphs, and then had humans review paragraphs to filter out false positives.

Computational and qualitative sociologists use different approaches to evaluating code reliability. We operationalize

reliability inclusively, as the overall level of agreement between coders (human or machine) at the end of the coding process. We use the $\alpha = 0.80$ intercoder agreement (ICA) our team reached using the “human”-only strategy as a reference (human-human ICA). For machine coding, we use F-1 scores, a data science metric that checks machine predictions against human coding for type I (false positives) and type II (false negatives) errors (human-machine ICA). We assess human effort by tracking the required number of hand-coded paragraphs for each strategy. This is represented as number of paragraphs coded, where higher values indicate more effort.

Figure 1 shows that the machine-centered strategy achieved acceptable reliability (F-1 = 0.63) while requiring human coding of 2,418 paragraphs to train BERT. Human-only coding of 9,673 paragraphs achieved an ICA of $\alpha = 0.80$. The “hybrid” strategy, which involved coding 2,418 paragraphs for BERT training and human review of an additional 1,471 paragraphs that the machine coded with less confidence, achieved the highest reliability (F-1 = 0.88) with a total of 3,889 paragraphs coded. In addition, the hybrid strategy rescued data points human coders missed (e.g., a paragraph with health information, “he checked it and then he wanted to take samples”). In analyses not shown, we obtained similar trade-offs between reliability and workload using other codes in our data set. Additional information and links to code are provided in the supplemental materials.

Our study shows that a hybrid approach that iteratively integrates machine learning into interpretative qualitative analysis can save hundreds of hours of human effort on a modestly sized project, while potentially improving coding reliability. It creates unique possibilities for scaling human coding to large volumes of primary data where human-only coding would be time or cost prohibitive and machine-centered coding would be inadequate. Yet we are not arguing that HHMLA should replace human in-depth interpretation in any qualitative approach. Rather, our goal is to provide resources and guidelines for sociologists interested in engaging with the computational social science (CSS) repertoire in ways that complement traditional qualitative analyses (Abramson et al. 2018). Researchers should assess the appropriateness, performance, and limits of these strategies given their own data, codes, research questions, and epistemic commitments. All research was reviewed and approved by the University of California, San Francisco institutional review board.

Acknowledgments

We would like to thank many friends and colleagues for useful feedback on prior versions of this work. In particular we would like to thank the members of the University of California, San Francisco, Medical Cultures Lab, the members of the University of Arizona Computational Ethnography Lab, and the New Mixed Methods Working Group for their feedback and support. We are particularly grateful to Alma Hernandez and Melissa Ma for their work and feedback on the broader project and Ron Breiger and Kelsey

Gonzales at the University of Arizona for their insightful comments and suggestions. We would also like to thank the editors and reviewers at *Socius* for helpful comments. Any remaining errors are our own.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: National Institutes of Health grant DP1AG069809 (Daniel Dohan, principal investigator), National Institutes of Health grant R01CA152195 (Daniel Dohan, principal investigator), and a Research, Discover, and Innovation Faculty Seed Grant at the University of Arizona (Corey M. Abramson, principal investigator). The opinions are those of the authors, not the funders, and any errors are our own.

ORCID iD

Zhuofan Li  <https://orcid.org/0000-0002-5104-528X>

Supplemental Material

Supplemental material for this article is available online.

References

- Abramson, Corey M., Jacqueline Joslyn, Katharine A. Rendle, Sarah B. Garrett, and Daniel Dohan. 2018. “The Promises of Computational Ethnography: Improving Transparency, Replicability, and Validity for Realist Approaches to Ethnographic Analysis.” *Ethnography* 19(2):254–84.
- Deterding, Nicole M., and Mary C. Waters. 2021. “Flexible Coding of In-Depth Interviews: A Twenty-First-Century Approach.” *Sociological Methods & Research* 50(2):708–39.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” *ArXiv*. Retrieved November 18, 2021. <https://arxiv.org/pdf/1810.04805.pdf>.
- DiMaggio, Paul. 2015. “Adapting Computational Text Analysis to Social Science (and Vice Versa).” *Big Data & Society* 2(2):1–5.
- Nelson, Laura K., Derek Burk, Marcel Knudsen, and Leslie McCall. 2021. “The Future of Coding: A Comparison of Hand-Coding and Three Types of Computer-Assisted Text Analysis Methods.” *Sociological Methods & Research* 50(1):202–37.

Author Biographies

Zhuofan Li is a PhD student in sociology at the University of Arizona. He studies the inter-institutional space between data-driven industries and data-driven science. His methodological interests include mixed-methods and computational analysis of texts and networks.

Daniel Dohan, PhD, is a professor of health policy, surgery, and humanities and social sciences at the University of California, San Francisco (UCSF). He is principal investigator at the Medical Cultures Lab, a scholarly community at the UCSF Philip R. Lee Institute for Health Policy Studies that works to understand medical culture and advance health equity.

Corey M. Abramson, PhD, is an associate professor of sociology at the University of Arizona. His research examines the connections between inequality, culture, and health. His comparative ethnography, *The End Game: How Inequality Shapes Our Final Years* (Harvard University Press), was released in paperback in 2017. His methodological works—including collaborative pieces published in

Sociological Methodology, *Ethnography*, and *Beyond the Case* (Oxford University Press, 2020, coedited with Neil Gong)—focus on (1) articulating the value of “ethnographic pluralism” and (2) developing strategies for integrating computational techniques to improve the scalability, transparency, and replicability of large qualitative and mixed-methods projects.