SenseMate: A Collaborative Human-AI Platform for Qualitative Coding

Keywords: Human-Machine Decision-Making, Natural Language Processing, Content Analysis, Qualitative Coding, Human-Computer Interaction

Extended Abstract

The growth of unstructured data allows researchers to more thoroughly study social phenomena. The data can be analyzed numerically or qualitatively through methods like content analysis. A crucial stage of content analysis is coding, which implies categorizing qualitative data to facilitate its interpretation. Structure is created by subdividing the data into units and assigning categories, or codes, to each unit. However, coding is tedious and time-consuming when carried out manually [1, 3]. An intuitive solution would be to leverage machine learning methods to fully automate qualitative coding. Though this approach would be efficient, it is prone to systematic errors or unfair biases that would negatively impact the downstream analysis of social phenomena. Finding a balance between manual and fully automated coding can help increase efficiency while allowing human judgment and preventing at-scale unfair machine bias. Several studies have interviewed researchers to identify opportunities for human-AI collaboration in qualitative coding and found that AI needs to be modifiable and transparent about their recommendations for successful collaboration [7, 4]. Previous work that created transparent algorithms for qualitative coding mainly applied keyword-based rules [7, 8]. We expand on these methods by applying a novel and interpretable machine learning strategy, known as rationale extraction models, to semi-automate qualitative coding. Rationale extraction models are integrated into SenseMate, an intentionally designed online platform, to facilitate bidirectional communication between AI and human sensemakers.

Data and methods. We are working with an annotated dataset containing 69 facilitated small-group conversations, in English, about people's experiences living in Boston, a United States city with over 650,000 people. These 69 conversations account for 175,899 words and 68 hours of transcribed audio. The community conversations are segmented into 1,170 snippets (a.k.a. units), which are responses to conversation prompts (e.g. what is your question about the future of Boston and your place in it?). We focus on a diverse subset of 9 themes (a.k.a. codes), which are enumerated in Table 1. The snippets are used to train rationale extraction models [5]. As Figure 1 depicts, rationale extraction models produce two outputs from a piece of text: recommendations for possible codes, or themes, and a corresponding rationale as to which words correspond to the recommended themes. We train rationale extraction models for each theme in Table 1, resulting in 9 separate models. We are actively retraining the models to incorporate user feedback on the suggested themes and rationales, so models can learn how users understand each code. To facilitate human-machine decision-making, we apply an iterative human-centered design process when creating SenseMate. After completing three design iterations, we are currently implementing a prototype of SenseMate, shown in Figure 2, which will be evaluated through a randomized controlled trial.

Results. Aiming to evaluate SenseMate as an entire system, we have started by separately assessing the model and design components of the platform. The designs have been evaluated through 13 user testing sessions, which are analyzed using affinity diagramming [6]. Regarding the model evaluation, information about the performance of each rationale extraction model is displayed in Table 2. It appears that more concrete themes, like "Race-based Inequality",

"Housing Affordability", and "Covid-19" have higher theme-detection accuracy scores compared to more abstract themes, such as "Processes" and "Community Values". In addition, we have run an experiment on Prolific (n=114), a crowdsourcing platform, to understand how the recommendations and rationales produced by the models impact qualitative coding. Participants are randomly assigned to one of three conditions while coding 13 snippets: (1) receiving no information from the models, (2) receiving only theme recommendations, and (3) receiving both theme recommendations and rationales. For each snippet, we record the selected themes, the coding time, and how confident participants feel about their answers. Figure 3 contains the main results from the experiment. We find that receiving assistance from the models had a significant effect on coding reliability and performance (e.g. accuracy, precision, and fscore) but not on efficiency or confidence. Participants with input from the rationale extraction models tend to perform better and obtain higher agreement compared to participants without access to the models. We may not have observed a reduction in coding time because participants were only asked to code a small number of fairly short snippets. An increase in efficiency may be observed when users interact with the entire SenseMate system to code hundreds of snippets.

Discussion. Our current results demonstrate that AI models can help human sensemakers code unstructured data by improving their coding accuracy and reliability. We are still exploring how the creation of human-interpretable explanations in addition to theme recommendations impacts coding performance within the entire SenseMate system. An important limitation, and current challenge, in our modeling approach is the small number of positive training examples we had access to. The growth of large language models, like GTP-3 and ChatGPT, offers promising avenues for future work on the modeling side of semi-automated content analysis, including the generation of training data. In sum, SenseMate is a working system that supports human-AI collaboration around data analysis. We hope this work sparks new ideas among computational social scientists on how to design and analyze new forms of human-machine decision-making.

References

[1] Tehmina Basit. "Manual or electronic? The role of coding in qualitative data analysis". In: *Educational research* 45.2 (2003), pp. 143–154.

[2] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *NAACL*. ACL, 2019.

[3] Annette Hoxtell. "Automation of qualitative content analysis: A proposal". In: *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*. Vol. 20. 3. 2019.

[4] Jialun Aaron Jiang et al. "Supporting serendipity: Opportunities and challenges for Human-AI Collaboration in qualitative analysis". In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1 (2021), pp. 1–23.

[5] Tao Lei, Regina Barzilay, and Tommi Jaakkola. "Rationalizing Neural Predictions". In: *Proceedings of the 2016 EMNLP*. Association for Computational Linguistics, 2016.

[6] Andrés Lucero. "Using affinity diagrams to evaluate interactive prototypes". In: *Human-Computer Interaction–INTERACT 2015: 15th IFIP TC 13 International Conference, Bamberg, Germany, September 14-18, 2015, Proceedings*. Springer. 2015, pp. 231–248.

[7] Megh Marathe and Kentaro Toyama. "Semi-automated coding for qualitative research: A user-centered inquiry and initial prototypes". In: *Proceedings of CHI Conference*. 2018.

[8] Tim Rietz and Alexander Maedche. "Cody: An AI-based system to semi-automate coding for qualitative research". In: *Proceedings of CHI Conference*. 2021.

Name of theme	**Definition**
Community Values	Values instilled throughout the community and differences in values within and across communities
Covid-19	COVID-19, vaccines, masks, COVID tests, boosters, and the impacts of COVID-19, such as working from home, school closures, and jobs lost
Housing Affordability	Cost of housing and how affordable that cost is to residents, regardless of tenure (tenant/owner) and subsidy (e.g. workforce housing, public housing)
Income	References to income/wages and wealth. This can include discussions about: one's personal income; satisfaction with their income; in/ability to increase their income; in/ability to build wealth; income inequality; the income/wage levels to be able to afford the cost of living in Boston
Processes	References to processes through which the public interfaces with government, such as voting, community engagement, campaigning, electoral processes, and other decision-making processes
Quality of Education	Education that empowers individuals and communities to get more control over their own situations and environments; education systems that focus on the importance of quality learners, quality learning environments, quality content, quality processes, and quality outcomes
Race-based Inequality	Defined as lack of jobs, services, and goods based on skin color, ethnicity, and language
Sense of Safety	Refers to feeling unsafe at home, in one's neighborhood, and throughout the city
Transportation	References to public transportation— like the MBTA, buses, and trains. This can include discussions about: the quality, affordability, accessibility, and safety of transportation

Table 1: A description of the 9 themes we focused on, including their names and definitions.

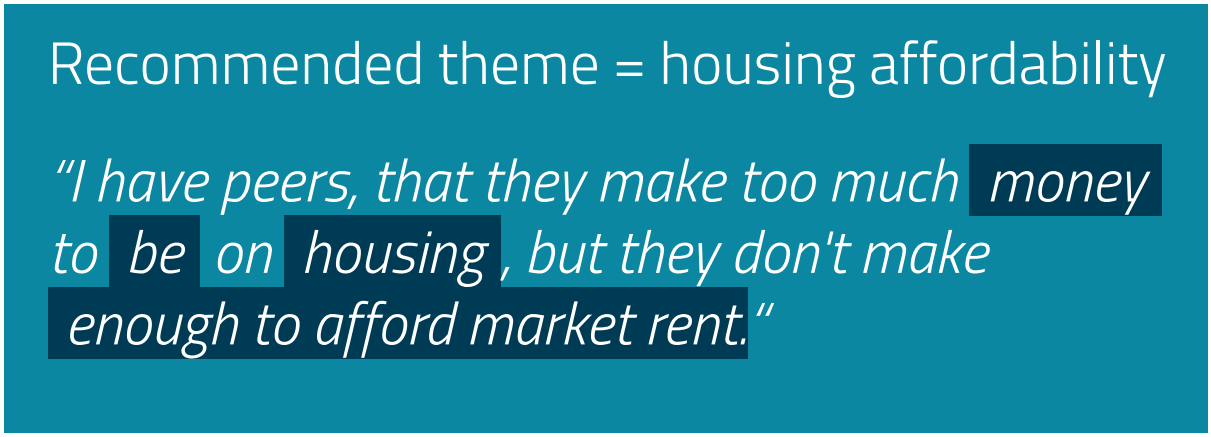
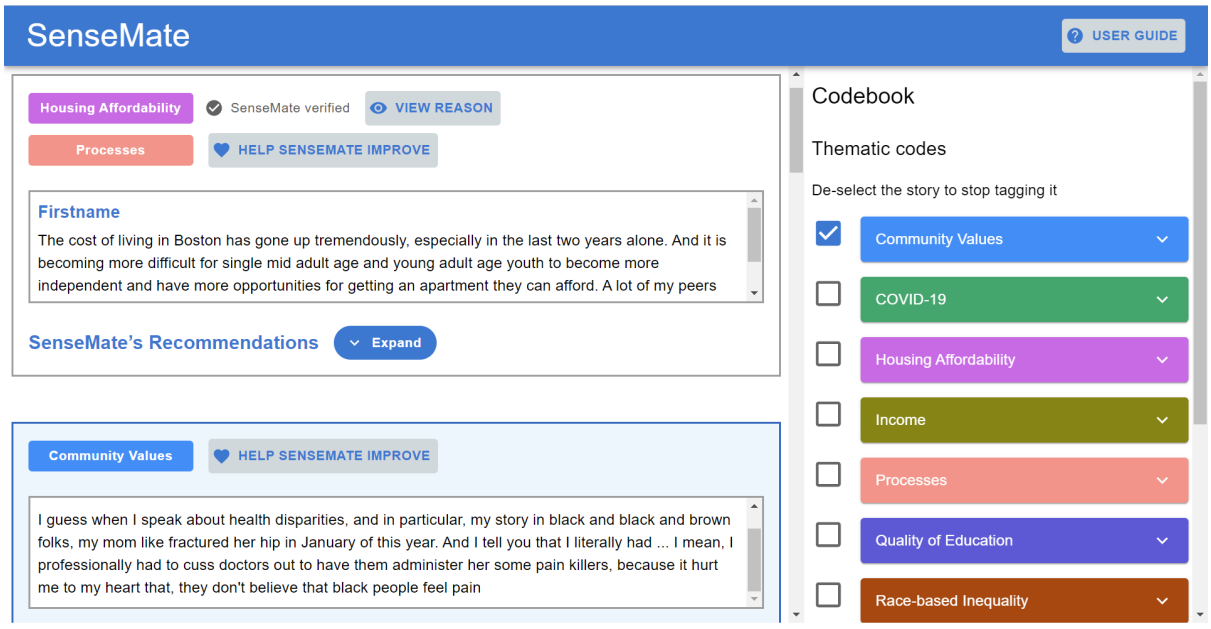
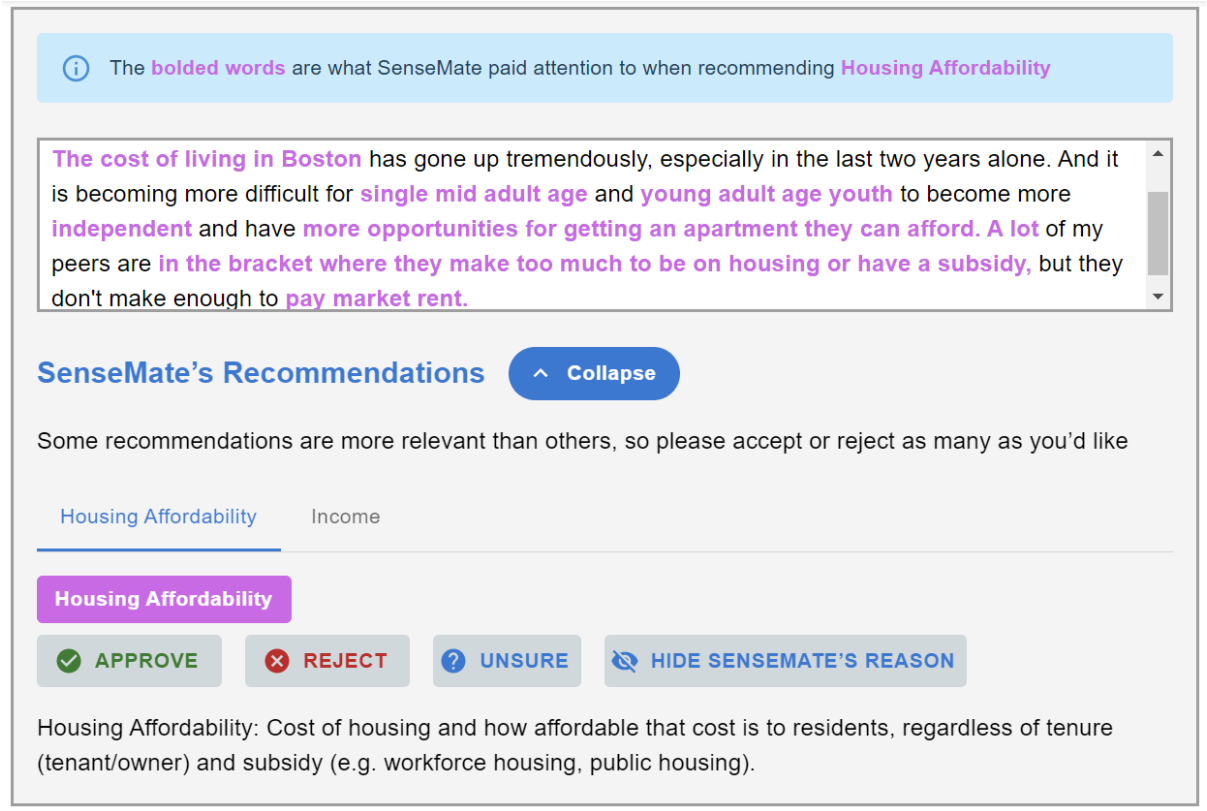


Figure 1: Rationale extraction models have two components: an encoder that classifies whether a conversation snippet contains a code or not and a generator that tries to identify the subset of words within a snippet that relate to the code (i.e. the rationale) [5]. Data is passed through the generator and then the encoder. Various model architectures can be used to create the encoder and generator. We are using the recurrent neural network (RNN) structures that were initially proposed by Lei, Barzilay, and Jaakkola [5] and augmenting them with BERT word embeddings [2]. In this example, the rationale extraction model correctly recommends the "Housing Affordability" theme. In addition, it outputs a rationale, which are the words highlighted in dark blue. The explanation shows how this sentence connects with housing affordability.



(a) The overall layout of the SenseMate platform. The left side shows a series of community stories that users would code. The right side shows the codebook that users select themes from. The first story has already been coded with the "Housing Affordability" and "Processes" themes. The second story is currently selected with the "Community Values" theme applied to it. Users can choose to expand the "SenseMate's Recommendations" section for each story to view and interact with recommendations from the rationale extraction models.

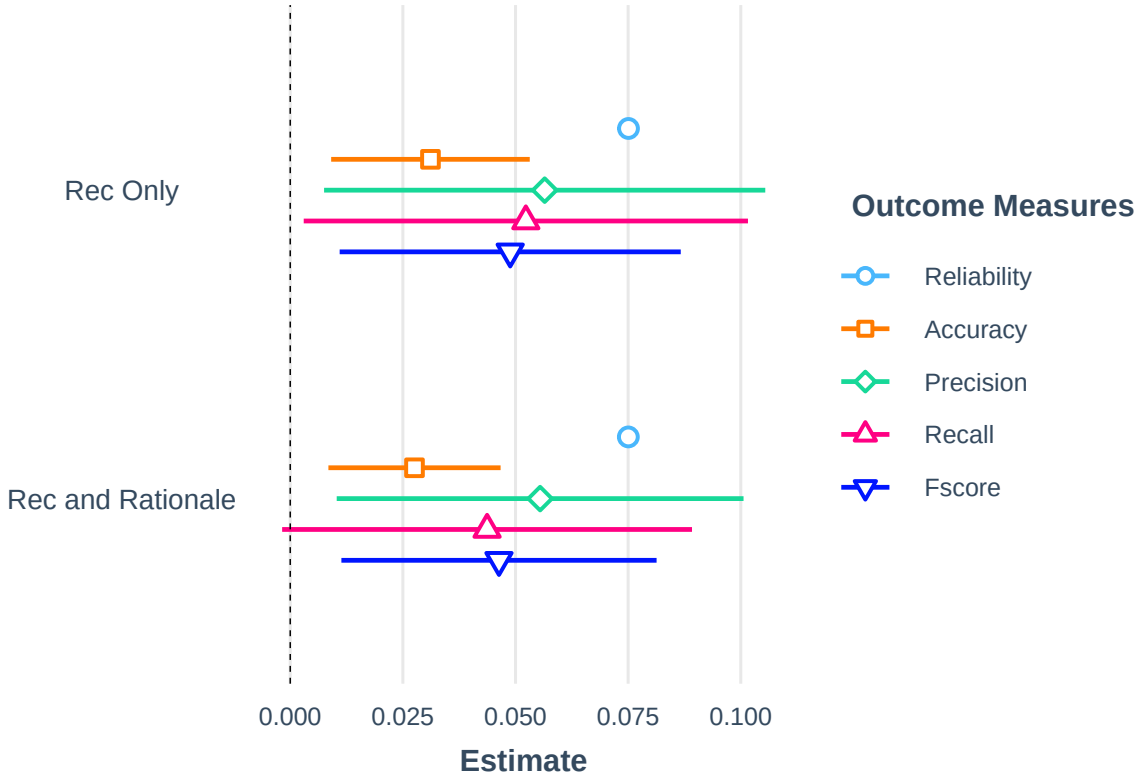


(b) A close-up of the "SenseMate's Recommendations" section. When users choose to expand SenseMate's recommendations, they can click through each recommendation. In this story, the "Housing Affordability" and "Income" themes are suggested. For each recommendation, users can complete the following actions: (1) approve the recommendation, (2) reject the recommendation, (3) mark the recommendation as unsure, or (4) view SenseMate's reason for the recommendation. This image displays what would happen if a user decided to view the reason for the "Housing Affordability" recommendation. The rationale generated from the models would be displayed in bold. From there, the user can approve or reject the recommendation, or even give feedback on the rationale.

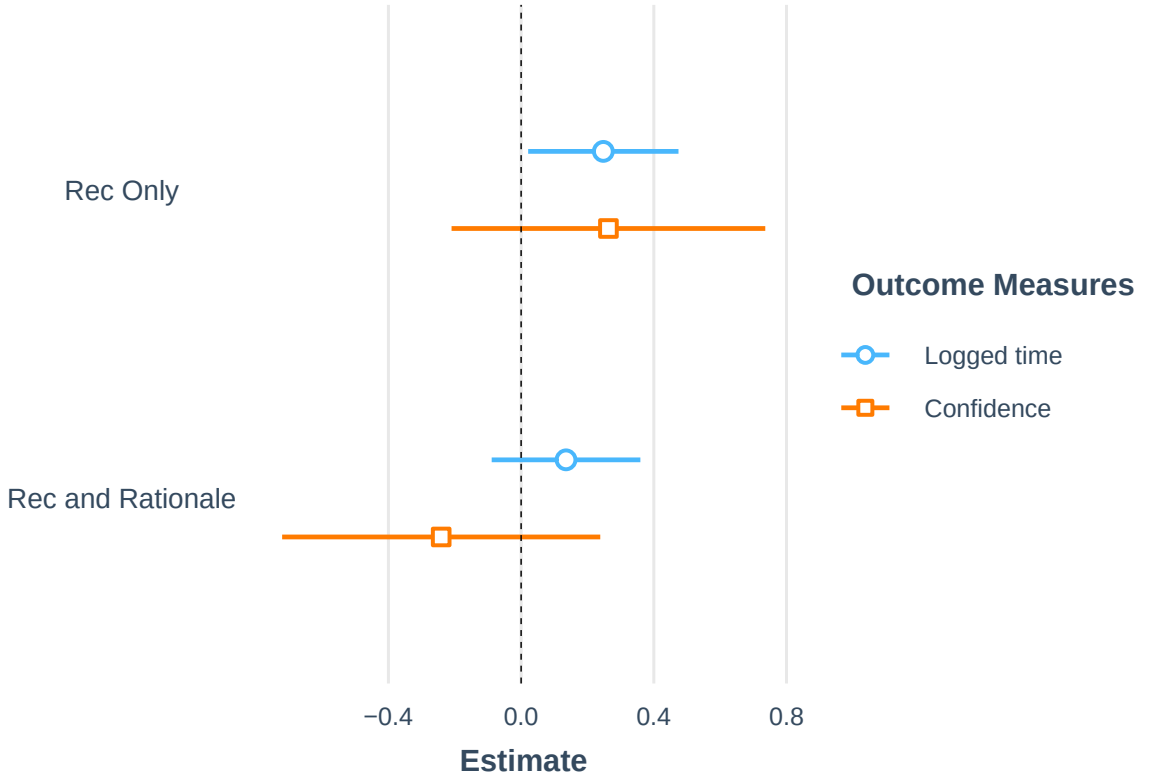
Figure 2: Screenshots of the SenseMate platform that is currently being built.

Theme	Accuracy	Precision	Recall	F-score
Community Values	81.3%	0.86	0.75	0.80
Covid-19	88.9%	0.94	0.83	0.88
Housing Affordability	90.9%	0.91	0.91	0.91
Income	84.4%	0.92	0.75	0.83
Processes	65.0%	0.71	0.50	0.59
Quality of Education	88.6%	0.90	0.86	0.88
Race-based Inequality	90.9%	0.93	0.89	0.91
Sense of Safety	84.4%	0.79	0.94	0.86
Transportation	87.5%	1.00	0.75	0.86

Table 2: A summary of the theme-detection performance of the rationale extraction models. We evaluate the performance of the rationale extraction models by comparing machine-generated recommendations with the human-labeled themes in the dataset. Predictions are binary, in which a positive prediction means a snippet contains a particular theme and a negative prediction means it does not. Model performance is determined by the following metrics: accuracy, precision, recall, and f-score.



(a)



(b)

Figure 3: Results from a Prolific study with 114 participants, 38 per condition. The control condition involves receiving no support from the rationale extraction models. The "Rec Only" condition involves receiving theme recommendations, and the "Rec and Rationale" condition involves receiving theme recommendations and corresponding rationales. Figures 3a and 3b show the associations between the treatment conditions and various outcome measures relative to the control condition. Circles represent average associations (i.e. regression coefficient values), and lines represent 95% confidence intervals. Intervals that do not intersect zero indicate a statistically significant association between the treatment condition and an outcome measure at an α level of 0.05.

In **Figure 3a**, the reliability outcome measure is obtained by calculating the Cohen's kappa between every unique pair of participants in each experimental condition. A linear regression model is created with coding reliability as the dependent variable and experiment condition as the independent variable. Linear mixed effect models are created for accuracy, precision, recall, and fscore, in which the experiment condition is the fixed effect and the user and snippet ids are random effects. All outcome measures except recall in the "Rec and Rationale" condition have positive and statistically significant coefficients. Positive estimates mean that participants in the treatment conditions have better coding performance and agreement compared to participants in the control condition.

In **Figure 3b**, linear mixed effect models are created for logged coding time and confidence. Coding confidence does not have statically significant coefficients, while coding time has significance for the "Rec Only" condition. A positive estimate means that the participants in the "Rec Only" condition take longer to code snippets compared to participants in the control condition.