An Approach to Computational Grounded Theory

Keywords: Computational grounded theory, Human-centred computing, Topic modelling, Grounded theory, and computational social science.

Extended Abstract

Introduction: Social scientists typically conduct research using qualitative methods, which require manual labour and are often impossible to apply to large datasets. In the age of social media, however, new valuable sources of social data offer researchers from different disciplines the opportunity to better comprehend various phenomena. We ,therefore, proposes a methodological framework for applying Computational Grounded Theory (CGT). The approach takes into account the depth of qualitative analysis and researchers' preferences to remain close to their analyses, rather than fully automating their interactions with the data. Using the proposed methodology, the aim is to explore tutors' experiences in the gig economy.
Data: Reddit was manually examined to find relevant subreddits using keywords and tutoring platform names, resulting in eighteen subreddits. To retrieve related discussions, we used the Reddit API Wrapper (PRAW), which resulted in approximately 52K posts and comments.
Methods: The framework consists of three phases - See figure 1: **Phase One: Data exploration-** The topics of tutors' discussions on Reddit were first identified in the initial data exploration step. Two randomly selected subreddits from our dataset were analysed using initial steps of the conventional GT coding process. This step aimed to gain early insights into the data and prevent the blind use of unsupervised machine learning algorithms. Subsequently, LDA Topic Modeling (TM) was used on the complete dataset with a threefold purpose: first, to ensure the validity and reliability of the codes identified in GT and vice versa; second, to extend the data exploration to include the entire corpus; finally, to extract the list of terms representing each topic produced by LDA and required to model the data in phase two.
Phase Two: Modelling - Query-Driven TM (QDTM) (Fang et al., 2021) was used to model the topics of the whole corpus. The model was fed with a set of query terms for each topic from the previses step, and it produced a list of main topics and subtopics.

Human evaluation of QDTM topics - Four tasks were presented to our annotators. First to judge the quality of the topics, then identify the issue with the once of lower quality then label the topics and finally judge the relatedness of subtopics to their main topics. Annotation guidelines were developed through several rounds of revision and testing within the research group. The annotation process was conducted in several stages.
Phase Three: Human-centred interpretation - Here we engaged with the conventional GT steps. Main topics are considered selective codes, while the subtopics are sub-codes. Topics' labels produced by the annotators are considered provisional. We begin with line-by-line coding of the 10 highest weighted posts of the first selective code in order to better understand, confirm, or modify the code, and then for each of its sub-codes (first group). This is an important step to help the researcher to apply the constant comparison process of GT and continuously stimulate ideas and writing analytical memos and defining links among the codes. Once the analysis of the first group has completed, we start analysing the next one until completing all groups. The process will result in constructing some categories that connect codes together. During the analysis we tried to conceptualise the patterns found in the data. Here, after deeply engaging with the data , the core category that represents the main concern can be define. Next, we engaged in theoretical sampling to saturate some of the categories. Then defining possible relationships between them, theoretical coding. This process is what helps to integrate the categories and data into a **Theory** that explains the main concern of participants and how they addressed or resolved it.

Results: Phase one - GT coding step resulted in 15 codes, then, two LDA models were collectively able to detect 12 topics identified in GT codes. However, the models failed to model one topic that was present in GT codes. Conversely, only one topic that was clearly modelled in both LDA models but not in GT. Since the aim of this step was to validate and further explore the data, and only one new topic was found even after the number of topics was increased, the aim of this step was fulfilled, and it was not necessary to examine further models. Thus, the final output of this step is 14 topics that will next be considered in QDTM. Noting that, two GT codes were excluded from the comparison with LDA as they were generated from observation of the data and due to their abstract nature LDA is not expected to model them.
Phase two: QDTM model 76 topics: 14 main topics and the number of subtopics for each main topic is automatically determined by the model. After the human evaluation step, topics (main topics or subtopics) that were assessed as incoherent, subtopics that were considered unrelated to their main topics, and topics that all three annotators disagreed on in one or more tasks were excluded. Lack of agreement in the issue identification task or the topic labelling task did not lead us to exclude these topics. For the annotation results (tasks 1,2 and 3) see Table 1. Consequently, 21 topics (27%) were excluded. Therefore we got 14 main topics and 41 subtopics. Data profiling of final topics see Table 2.
Phase three: The theory - In short: The main concern for tutors in the gig economy is **staying afloat**. To process and resolve their main concern, they engage in a series of behaviours. First, create a supportive context: Component of this context are novices the help seekers- in the context of confusion, and experts the help providers- in the context of clarity. After this interactions, a state of realization is reached among novices. Therefore, novices, as well as experts, in the context of clarity, are divided into two main positions: either refuse how these platform work then leave or accept and try to solve. To solve, tutors use number of strategies such as building and maintaining base of regular students, protecting reputation, and work on multiple platforms, depending on their level of competency.
Conclusion- The use of computational techniques provided us with a manageable amount of representative data to analyse and significantly sped up the analysis. In particular, it reduced the amount of data from 52K posts to 550 posts. Furthermore, QDTM, a hierarchical structured TM, was found to be better suited to conducting CGT than a single-layer TM (the latter is used by existing CGT approaches such as Nelson (2020) and Odacioglu et al. (2022)) since it presented the researcher with groups of related topics (codes) rather than a large number of codes with no discernible relationships among them which facilitate the constant comparison process. By analysing these related codes together, the researcher will be able to uncover relationships among them more quickly and thus develop a deeper understanding of them. Moreover, the inclusion of human evaluation has proven extremely useful in ensuring the quality of the topics as posts on the same topic have been judged to be related and to discuss similar issues which have resulted in the researcher only presented with highly coherent topics to analyse, this saves time and effort. This step was also not included in previous approaches.

References

Fang, Zheng, He, Yulan & Procter, Rob. 2021. A Query-Driven Topic Model. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Online. Association for Computational Linguistics, pages 1764-1777. https://doi.org/10.18653/v1/2021.findings-acl.154

Nelson, Laura K. 2020. Computational grounded theory: A methodological framework. *Sociological Methods & Research,* 49(1)**,** 3-42.

Odacioglu, Eyyub Can, Zhang, Lihong & Allmendinger, Richard. 2022. Combining Topic Modeling with Grounded Theory: Case Studies of Project Collaboration. *arXiv preprint arXiv:2207.02212*.

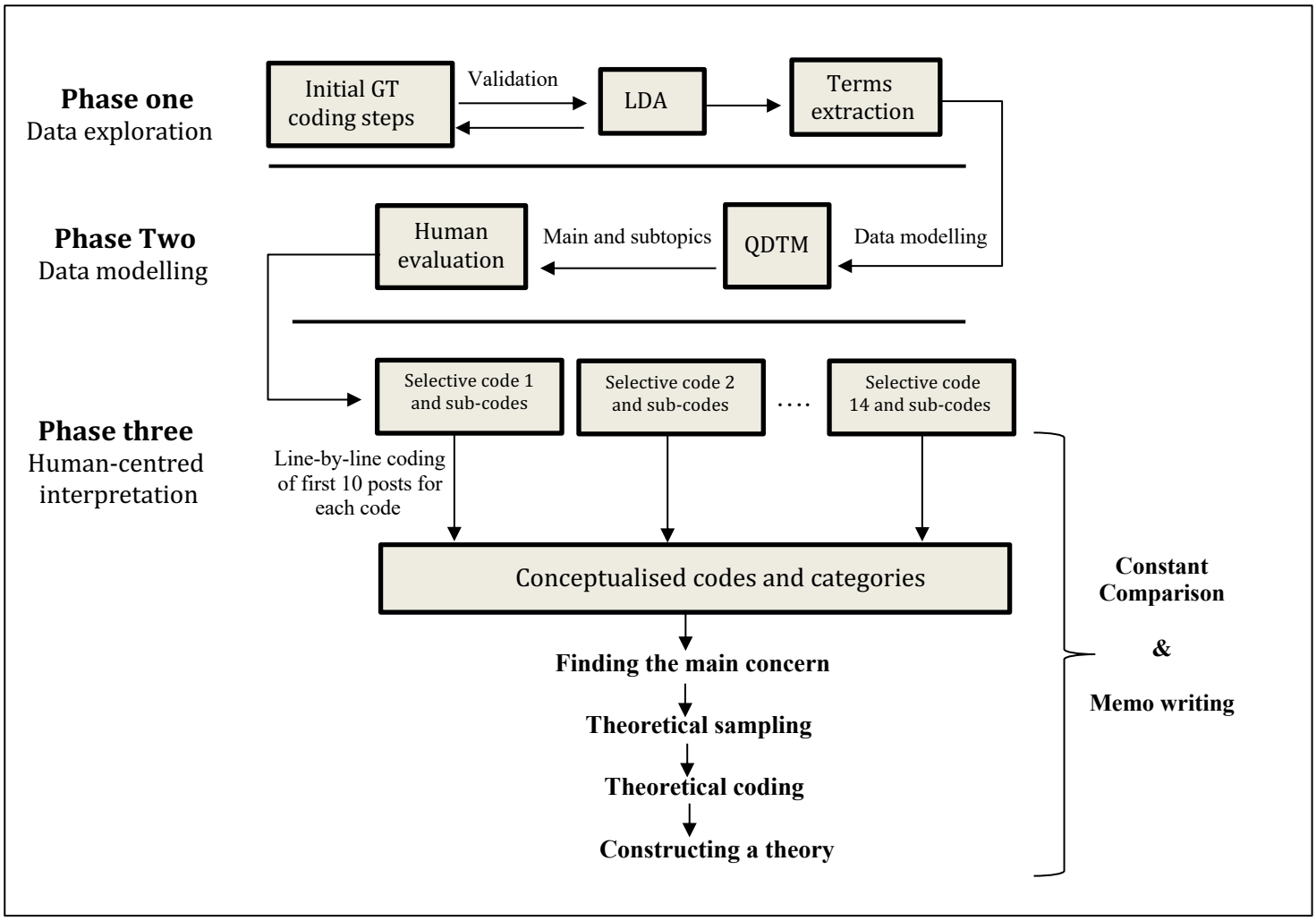

Figure 1: Methodology chart

Stage	Task	All agree	Two agree	No agreement	Total number of topics	Fleiss' kappa
One	1- Topic coherence	6 (50%)	6 (50%)	0	12	0.34
	2- Issue Identification	5 (42%)	7 (58%)	0	12	0.38
	4- Relatedness to main topic	4 (40%)	6 (60%)	0	10 (only applied to subtopics)	0.38
Two	1- Topic coherence	32(42%)	43 (57%)	1 (1%)	76	0.33
	2- Issue Identification	24(32%)	49 (64%)	3 (4%)	76	0.35
	4- Relatedness to main topic	18 (29%)	41 (66%)	3 (5%)	62 (only applied to subtopics)	0.21

Table 1: Breakdown of annotators' decision counts for tasks 1,2, and 4 and Fleiss' kappa values.

Topic number	Number of subtopics	Prevalence of the topic across the corpus	Labels of main topics that at least two annotators' agreed on *
1	3	2.98%	Recruitment process/ Recruitment
2	5	6.35%	Qualifications/ Language qualifications
3	2	14.67%	Teaching levels/ Teaching
4	5	16.16%	Booking rates/ Attracting bookings/Bookings
5	4	4.14%	Pay/ Pay rates
6	1	8.44%	Ratings/ Rating system
7	4	5.48%	Internet/ IT problem
8	3	1.73%	Bank transfers/Exchange
9	1	5.68%	New contracts/ Contracts
10	2	6.15%	Pay conditions/ Income
11	7	6.82%	Lessons/classes
12	0	4.93%	Business structure/ employee relation
13	2	2.68%	Technical problem/ Issue
14	2	3.65%	Online teaching/ Teaching opportunities

Table 2:Number of subtopics per topic and topic prevalence.

* For labels agreement: A python script was written to automate this process using the FuzzyWuzzy and Wordnet python libraries to find matching and synonymous labels.