Establishing Trustworthiness Through Algorithmic Approaches to Qualitative Research

Ha Nguyen(✉), June Ahn, Ashley Belgrave, Jiwon Lee, Lora Cawelti, Ha Eun Kim, Yenda Prado, Rossella Santagata, and Adriana Villavicencio

University of California-Irvine, Irvine, USA
thicn@uci.edu

Abstract. Establishing trustworthiness is a fundamental component of qualitative research. In the following paper, we document how combining natural language processing (NLP), with human analysis by researchers, can help analysts develop insights from qualitative data and establish trustworthiness for the analysis process. We document the affordances of such an approach to strengthen three specific aspects of trustworthiness in qualitative research: credibility, dependability, and confirmability. We illustrate this workflow and shed light on its implications for trustworthiness from our own, recent research study of educators' experiences with the 2020 COVID-19 pandemic; a context that compelled our research team to analyze our data efficiently to best aid the community, but also establish rigor and trustworthiness of our process.

Keywords: Natural language processing · Trustworthiness · Qualitative research

1 Introduction

> "We need to really study what other districts are doing and and how it's working, how it's working at their district. So working collaboratively with other school leaders at other districts …." – [Principal, Local Elementary School, Interview Transcript].

The global, COVID-19 pandemic in 2020 brought about tremendous distress and disruption in social life, economic systems, and civic institutions such as schooling and government (and as of this writing, continues to threaten societies around the world). As a team of education researchers in the United States (U.S.), we faced a unique situation of wanting to collect data on how educators, parents, and families were experiencing this crisis while also providing rapid information to inform the questions that our local educator partners were facing. The consequences of our research in this time were substantial. Key decisions such as how to bring students back to physical campuses, or support learners and their families in remote learning, had dire repercussions for public well-being.

Within this context, we conducted in-depth interviews with educators, school staff, and parents in California school districts. Many interviewees articulated the need for fast

A. R. Ruis and S. B. Lee (Eds.): ICQE 2021, CCIS 1312, pp. 47–61, 2021.
https://doi.org/10.1007/978-3-030-67788-6_4

turnaround of research analysis to help them understand a wide variety of topics, from how to differentiate instruction to addressing pandemic-induced community trauma. These needs brought us to consider computational approaches for analyzing a rapidly evolving qualitative data corpus. This experience of utilizing computational approaches to aid our analysis, heightened by the context of the pandemic, brought enduring issues in qualitative research into stark relief. Qualitative research is typically suited for *slow scholarship*, where an analyst can take their time to explore data, and undertake a rigorous process to find relationships and develop deeper, theoretical understandings of a phenomenon.

Instead, to address the fast timelines needed by educators who would use our research, we sought an algorithmic approach by utilizing natural language processing (NLP) to provide an initial parsing, categorizing, and clustering of our qualitative data. Then, we introduced *humans in the loop*, as our research team worked with the outputs of different NLP approaches to analyze and synthesize potential insights. We were inspired by prior frameworks for bringing computation into both deductive (i.e., theory-grounded) and inductive (i.e., data-grounded) analysis workflows [2, 3, 20].

This paper documents what these frameworks may look like in educational research contexts for analysts conducting the research. In the process, we contribute to the emergent research on algorithmic analyses in qualitative work in two main ways. First, we illustrate how different, NLP approaches to parsing, organizing, and presenting raw qualitative data substantially influence the sensemaking and research directions that researchers may take. Understanding how this mutual influence between computation and human insight may intersect, is important to map out methodological transparency in future studies.

Second, our analyses illuminated how combining NLP and algorithmic approaches has potentially major implications for establishing three specific facets of trustworthiness: **credibility, dependability, and confirmability**, which are fundamental aspects of qualitative methods. We describe a set of guidelines for reporting research processes to establish trustworthiness; particularly to map the relationship between data sources, algorithmic choices, development of data patterns, the human sensemaking process, and triangulation of information from both computer and human analysis. This discussion is vital because how researchers make public all facets of their research, is key to establishing rigor in qualitative analyses [1].

2 Theoretical Framework

2.1 Algorithmic Approaches to Qualitative Research

Researchers across fields – such as digital humanities, psychology, communication studies, and education – have applied computational techniques to uncover insights through analysis of texts [4, 19, 25]. A common workflow is to use natural language processing (NLP) to extract aggregate counts of parts of speech (e.g., pronouns, nouns, verbs), word usage, and topics, and map them to predefined categories [11, 25]. The mapping of words to categories is often grounded in prior theories and researcher assumptions.

Researchers can come in with preexisting libraries or code categories [23, 25]. Alternatively, researchers can read through several manuscripts, directly refine the code categories, and search for the exact words or phrases in the data corpus [19]. An advantage of using keyword matches is that researchers can automate the analysis and establish inter-rater reliability between the researchers and the classifiers [6].

A limitation to finding exact matches for pre-defined codes, however, is that this approach requires intensive researcher labor to carefully examine raw qualitative data, establish themes, and validate new categories. In rethinking this workflow, researchers have suggested analytic pipelines to automatically search for words with related meanings; for example, using NLP to conduct a search of keywords based on semantic similarity [3]. Researchers in this area suggest use of contextual word embedding to create word vectors from the contexts of the vocabulary (i.e., based on its relation to the surrounding words) and search for semantically similar words based on the contexts.

Researchers have also proposed computational techniques for bottom-up analyses, where keywords and themes can emerge from data without relying on predefined categories. Emergent work has suggested the intersection of computation and ethnography [20]. Both unsupervised machine learning and qualitative methods – such as grounded theory – are inductive and driven by the data [8]. Both approaches value the importance of data contexts in informing interpretation [20]. For example, machine learning researchers have emphasized the role of *human experts in the loop*: while computation can identify semantic patterns, the meaning-making of these patterns depends on human judgment [9].

There is emergent education research that uses computation in inductive analyses. Traditionally, qualitative researchers conduct content analyses by reading the text manually and identifying themes and categories. Recently, researchers have proposed use of topic modeling to discover latent topics in education data corpus [2, 14]. Although topic modeling has shown potential in uncovering broad-based themes, researchers maintain that there is a need for human's domain knowledge to define more fine-grained topics [2]. However, there is limited research that details the analysis workflow for combining human sensemaking and computational outputs, or discusses the implications of presenting these workflows to the public to establish trustworthiness in data interpretation.

2.2 Trustworthiness, A Key Element of Qualitative Research

One critique of qualitative research is that researchers often fail to provide clear justifications for their study designs, analyses trails, interpretations, and claims [7]. In response, qualitative methodologists have suggested several standards for evaluating quality and rigor. For example, Lincoln and Guba [15] translate validity criteria found in quantitative work – internal validity, external validity, reliability, and objectivity – to facets of **trustworthiness**. In place of internal validity (i.e., the extent to which researchers can infer a relationship between variables), qualitative researchers probe for **credibility**, or the extent to which the data and interpretations are plausible and accurate. To establish credibility, researchers can make use of strategies such as triangulation among multiple researchers and data sources, member checks of one's interpretations with participants, and constant comparisons of emerging patterns and data. In place of internal reliability

(i.e., the extent to which data collection and analyses are consistent), researchers establish **dependability** by communicating the consistency of the research process. Potential practices to address dependability include use of inter-rater reliability, strong and logical mapping of study designs to research questions, and multiple checks of analysis between researchers. Finally, while research in positivist and post-positivist traditions seek to establish objectivity, qualitative researchers who work in more interpretive traditions, instead strive to establish evidence of **confirmability** in their studies; or "audit trails" of how analyses can be traced back to original data sources.

Despite these guidelines, "there is a lack of will and/or means" to make public the research collection and analysis processes in many qualitative studies [1] (p. 29). If researchers do not spend time explaining how the themes in their findings emerged, readers may have difficulties verifying whether the findings bear congruence with the actual phenomenon under investigation [7]. Even though researchers frequently mention triangulation and member checks to establish credibility, they may not detail how these processes are achieved. Consequently, Anfara et al. [1] propose that researchers need to clearly document the iterations of study design and analysis. Examples of these forms of documentation are mappings of interview protocols to research questions and mappings of emergent themes from analyses to initial codes grounded in data [1, 17].

In the following paper, we argue that computational approaches **can help sharpen notions of trustworthiness in conducting qualitative research**. For example, computational outputs play a role in developing insights into the major themes and sentiments in the text [20]. One might also conceptualize computational algorithms as an external coder, in collaboration with a team of human researchers, and as such might provide evidence to establish credibility and dependability [4, 6, 25]. By establishing practices of explaining different algorithms, and their potential influence on the qualitative sense-making process, we argue that researchers might also strengthen the confirmability of their studies. In essence, aligning computational models, the data used, and human interpretation–a process known as *closing the interpretive loop* [24]–is key to checking the validity of the model and its interpretation for researchers.

To illuminate this link between algorithms and qualitative analysis, and its implications for strengthening trustworthiness in the analytic process, we present a self-study narrative of the initial stages of a recent, research endeavor [5]. As a self-study, we note that we are not presenting a traditional research study (with the expected paper sections such as methodology, findings etc.). Instead, our documentation represents a layer of *meta-awareness and reflection* of our methodological process itself. The self-narrative we share serves to shed light on how computational approaches can strengthen qualitative research, as we explore the following question:

> What are the affordances of algorithmic approaches for developing and examining research directions, analyses, and consistency of findings?

3 Setting the Study Context

Our methodological insights derived from a research study that occurred in Spring-Summer 2020. Our research team was situated in an educational partnership between university researchers, schools, and school districts in California, U.S. To support our

partners to prepare plans for schooling and supporting students and families in a time of crisis, throughout May and June 2020, we conducted in-depth interviews with 35 district administrators, principals, teachers, school staff, and parents from our network. Each interview lasted approximately 45 min. The need for efficient turnaround of research led us to consider the potential of computational analyses in highlighting key data patterns.

In our initial phase of analysis, we sought to examine the feasibility of computational approaches in parsing the raw interview transcripts and aiding the research team in identifying areas to focus our analysis. We selected a subset of the interviews with school and district administrators (10 interviews; 48,567 words). Interviewees came from an array of instructional contexts (e.g., elementary, middle, high school; public and charter).

We then applied different NLP algorithms to provide an initial parsing, categorizing, and clustering of our interview corpus. Our research team in this phase, included the co-authors of this paper (6 PhD students in Education, and 3 faculty members who served as the principal investigators of the project). All members of the research team were involved in conceptualizing the research study, as well as recruiting and conducting interviews. Thus, everyone on the research team had prior knowledge that they brought to the analysis process.

In the next sections, we outline the steps we undertook to categorize, cluster, and interpret the data. As we progress through this self-study narrative, we illuminate the key implications for qualitative methodology that emerged and became clear in different stages of the process: selecting and running algorithms, engaging the human researchers in the analytic loops, checking for inter-rater reliability, and triangulating findings.

4 Algorithmic Transparency as a Step Toward Trustworthiness

The first insight we derived from our process can be described as follows:

> Communicating key information about our algorithm choices is vital for understanding how insights and findings are ultimately derived from the research process. The design and implications of algorithms provide evidence for stronger **credibility, dependability and confirmability**.

We illustrate this insight by describing our use of two NLP approaches on the same data corpus: (1) a deductive approach, where codes are generated from the interview design and word clusters are automatically identified based on keyword similarity; (2) and an inductive approach using a pre-trained text model to create topic clusters without researcher keywords.

4.1 Deductive Approach

Word Embedding. The overarching idea of our *deductive* approach is to first identify keywords based on *researcher input* (i.e., the interview questions in our research protocol). For example, our interview questions asked about remote learning, supporting students, equity issues during the pandemic etc. We then parsed the data to select the words that co-occurred with the keywords in similar contexts. We used Word2Vec [18]

to train word vectors (representation of words as feature vector) based on the data contexts. The learning model used a Continuous Bag-of-Words (CBOW) approach, which created the embedding by predicting the target word based on its surrounding words. This approach is based on the Distributional Hypothesis that words that appear in similar contexts are likely to have related meanings [13]. In particular, the local contexts of the words were defined by a sliding window of its neighboring words. Consider an example sentence: "Families can pick up meals at the schools around noon". In our case of a sliding window of size 5, the context for the word "meal" in the example was created using the 5 words before ("families", "can", "pick", "up") and the 5 words after the target word ("at", "the", "schools", "around", "noon"). The size of the sliding window influences the vector similarities: smaller windows produce more syntactic closeness, while large windows (e.g., commonly of size 5) generate broader topical groupings [12].

Keyword Search through Semantic Similarity. We then created word clusters by identifying the most similar words to a list of a priori keywords. The keywords were picked by us from the themes in our interview questions. Our keywords covered the following topics: technology access (e.g., "technology access", "devices"), approaches to distance learning (e.g., "distance learning", "online learning"), parental responses (e.g., "parents", "challenging families"), teacher collaboration (e.g., "teachers collaborate"), district policies (e.g., "district policies"), and responses to vulnerable populations (e.g.,"ela", "homeless").

4.2 Inductive Approach

Part-of-Speech Tagging and Word Embedding. We also analyzed our interview corpus using an *inductive, algorithmic* approach. The strategy of our inductive coding is to identify the noun phrases from the interview corpus, and cluster these phrases into topical groupings.

Table 1. Example output from deductive

Keywords	Words
School closure	Facebook, school sites, providing, wifi, ap, six weeks, two weeks, packets, instruction, decisions, game, ideas, open, were trying, small group, email
Food insecurity	Businesses, relevant, successful, complaints, member, income, tried best, dilemma, gift, guy, separated, unprecedented grader, laptops, vulnerable students, administrator, collaboration, finish, we talked, impose, yelling
Distance learning	Face, put together, model, local, brick mortar, feel, eight, scheduling, thursday, mental health, terms, person, grading, devices, make sure, work home, world
Teachers collaborate	Before, days, expectations, daily, create, activity, virtual, grading, students learning, brick mortar, try, staff, pandemic, social, deliver, translate, facebook, job, mandated, ed services, problem
School district	Policy, pd, problem, make, virtually, promotion, scheduled, level, daily, phone, bring, meetings, virtual, packets, creating, enrichment, speakers, offering

We obtained the noun phrases from each sentence of the interview corpus through part-of-speech tagging (POS). POS classifies a word in a text as corresponding to a noun, adjective, or adverb, etc., based on its definition and its adjacent words. We then used Python spaCy's dependency parser to identify the nouns and define their modifiers (e.g., adjective, adverb, or another noun). For example, in the example sentence "I prefer breakfast food", "food" will be identified as the noun, and "breakfast" as the modifier. In total, we identified 1833 unique noun phrases from the corpus.

Upon identifying the phrases, we used the large pre-trained model from spaCy to create a word embedding model for the phrases. The model contains 300 dimensional word vectors that were trained on a vocabulary of 2 million words from web page data (Common Crawl dataset; GloVe, [21]). We then worked to cluster words together using the word embedding developed with spaCy.

Table 2. Example output from inductive

Group	Words
1	Worksheets activities, pamphlets resources, resource guide, family response, stick Chromebook, follow guidelines
2	**Distance learning**, mastery learning, emergency learning
3	Roofs head, avenue support, corporations part, textbook students, taxpayers money, terms work, terms food, terms collaboration, hotel connectivity, family access, kids opportunity, brick mortar
4	Grading policy, learning format, learning school, pacing guide, teaching learning, giving vision, address learning, supervisor vision
5	**School closures**, school districts, school sites, phone calls, budget cuts, work students, contact students, support families, disinfect schools, respond students, check ins, vento liaison, school program, school closing, school community, school homework, summer school, counseling school, community facilitator, lunch community, school model, school board, school level, district office, district sign, partnership district

Notes. Bold text highlights the keywords from the deductive coding

Clustering. We created word clusters based on the Cosine similarity between the word vectors, using DBSCAN (Density-Based Spatial Clustering of Applications with Noise; [10]). DBSCAN is an unsupervised clustering method that has been used with high-dimensional data such as text vectors [26]. The algorithm was chosen because it did not enforce that all samples group into a certain cluster (i.e., allowing for noise and single-word cluster outliers).

The algorithm worked as follows: DBSCAN first divided the data set into n dimensions and formed an n dimensional shape for each data point. DBSCAN iteratively refined the clusters by going through each data point to determine if (1) the distance between points was within a user-specified radius (i.e., Eps) and (2) the clusters met the predetermined minimum number of points (i.e., MinPts). To create a manageable workload of word clusters, we created the clusters from the most frequent 500 noun phrases in the dataset.

The clustering generated 26 clusters, whose word counts ranged from 3 to 27 (Eps = .11, MinPts = 3). To select the optimal value for Eps, we followed the procedures outlined in [22]. We calculated the distance between a data point and its nearest neighboring points, plotted the distance in a k-dist plot, and selected the point of maximum curvature in the plot as the Eps value.

4.3 Algorithm Outputs and its Contributions to Establishing Trustworthiness

The outputs from the two computational approaches were quite different (see Table 1 and 2). Take an example of the phrase "distance learning". Results from deductive analyses yielded terms such as "face", "scheduling", "devices", "grading", and "mental health". Although "distance learning" appeared in a cluster in the inductive codes, this cluster contained a different set of words, grouped together with the terms "emergency learning" and "mastery learning".

Our documentation of the inner workings of our algorithm choices, along with the influence of the parameters on the final output, have implications for establishing credibility, dependability, and confirmability.

Credibility. Being transparent about the algorithm's functionality provides a form of member check to allow the research team (and readers) to understand the interpretation process. Interestingly, we are implicitly treating an algorithm as "another set of eyes" and almost like another member of the research team. Credibility is strengthened if readers can hopefully examine whether the interpretations that emerge from the data and computational output are plausible, given the internal design and implications of a given algorithmic approach.

Dependability. Some ways to strengthen the dependability argument in a qualitative study is to record the methodological and interpretive process of the researchers. We observe that an algorithm – in essence – is acting as another coder.

Thus, being transparent about the algorithm's process is a key element of establishing dependability. In addition, one affordance of algorithmic approaches could be in establishing criteria such as inter-rater reliability (IRR) between researchers. We delve into the opportunities for considering IRR in more detail, below, when we bring the human in the analysis loop.

Confirmability. One common strategy for establishing confirmability in qualitative research is to provide readers with an "audit trail" of steps and process that a research team undertook. This audit trail, when successful and strong, allows other scholars to follow the process and evaluate its logic, match to research aims, and links to findings. Here we note that transparently explaining algorithmic approaches in the process provides a synergistic way to describe an audit trail, and even potentially strengthen ways for other researchers to trace the analysis steps and replicate the interpretation.

5 The Human-Computer Analytic Loop: Strengthening the Trustworthiness Argument

> Any given algorithmic output substantially influences the analysis directions of the research team.

We illustrate this insight by describing the process of engaging the entire research team to continue with the analysis process, now aided by an initial look from different NLP algorithms, as a first pass to parse and organize the data.

We randomly divided ourselves into two groups, with each group focusing on understanding the word clusters that were derived by the two algorithms. The first two authors, who took the lead in developing the algorithmic process, also facilitated and took observation notes of the analysis groups. We observed that two sense making activities were occurring as our groups continued to analyze the word clusters: deriving new questions that could focus analyses and developing conjectures for what might be happening in the interview transcripts.

5.1 Algorithmic Output Influences Researcher Sense Making

We observed that the research team readily built conjectures around the data when encountering the word lists. Team members noted that the clusters helped to reduce their cognitive load, as "all the words were on screen" [Member 2]. We attempted to make sense of the word clusters based on their face values as well as prior experiences in educational settings. Member 4 reflected:

> I came up with the themes not purely from the grouping [of words] but because of the added knowledge and experience about each word that emerged. I had to make sense of 2 or more words together and draw on my experience to then come up with the theme. Other themes probably exist that I cannot see because I haven't had related experiences.

Relying on the different word clusters that came about from different algorithm choices resulted in varying conjectures and questions. The key insight here is that the analysis direction that a research team goes down can be substantially influenced by the algorithm choices that are made.

For example, when looking at a word cluster related to "distance learning" from the inductive approach (Table 2), members of our team observed that the relations between "distance learning", "mastery learning", and "emergency learning" could depend on the school's infrastructure to support these different types of learning. Another member voiced that maybe interviewees were talking broadly about different learning models. Another researcher proceeded to propose research questions from this inductive approach: "What types of learning responses are being offered? How do responses vary with different school representations in the dataset?" We note that these types of questions lean towards categorizing responses in a cross-sectional way, using broader descriptors.

Meanwhile, when observing the word clusters that came from a deductive, algorithmic approach, the words associated with "distance learning" included: face, put together,

model, brick and mortar, scheduling, mental health, grading, devices etc. Some research team members conjectured that the interviews might reveal the challenges in transitioning to this form of learning. Other members observed that the word clustering here, oriented our thinking to asking the "how" questions, or the details of implementing distance learning.

5.2 Strengthening Credibility and Confirmability

This vignette of how our research team interacted with the algorithmic output have implications for establishing credibility and confirmability.

Credibility. Using algorithmic approaches in the initial analysis phases, may provide qualitative researchers with an additional building block to establish credibility arguments. One simple way to illustrate this affordance is by explaining, as we did above how different algorithmic approaches connected to the varying, iterative interpretive processes of sense-making our research team members undertook.

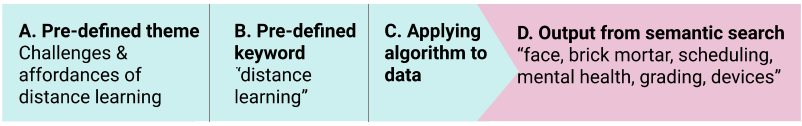

Fig. 1. Code Mappings for "Distance Learning", Deductive Approach. Blue (Steps A,B,C): Human; Pink (Step D): Computer. (Color figure online)

Tables 1, 2 and Fig. 1 serve as one potential example for how to align researchers' assumptions, data computational output, and interpretative cycles to identify emergent patterns for analyses. We note that researchers should clearly denote where in the process computational analyses come in to shape their assumptions around data and show whether the iterations of interpretations between the researchers and the algorithms are believable and logical. Furthermore, to establish transparency in mapping research questions and analyses, researchers can preregister the expected data outputs to justify why they select certain computational and interpretative strategies. Lastly, we were intrigued by the different conjectures that arose from merely looking at the algorithmic output. As our team delves back into the actual interview transcripts, documenting how subsequent interpretations related to, confirmed, or ran counter to the conjectures created from algorithmic output would greatly enhance credibility.

Confirmability. A key strategy to establish confirmability in qualitative work is to present an "audit trail", explicitly mapping how original data sources link to subsequent research choices. We note here that algorithms and their output represent another source, and a clearly documented trail of how decisions linked back to raw interview transcripts and algorithmic decisions, in an iterative way. In addition, a potential affordance of combining algorithmic approaches with qualitative reporting is to provide a roadmap for other scholars to follow through the steps of interpretive analysis. Fig. 1 provides an example roadmap, where other scholars can clearly trace our steps from pre-defined themes (Fig. 1.A) and keywords (Fig. 1.B) to the computational output (Fig. 1.D).

6 Algorithmic Affordances for Inter-rater Reliability

Communicating the nuances in refining automated classifiers is key to establishing dependability–the extent to which research processes are consistent and reliable.

To establish dependability, qualitative researchers examine inter-rater reliability (IRR) in their coding process. We observed that the codes and their associated keywords from our algorithms (Tables 1 and 2) could be used as keywords for an automated classifier, which can then act as a second coder in collaboration with a researcher. The classifier would identify whether a code is present in the text based on the occurrence of the keywords. Emergent research has suggested the potential of automated classifiers to establish IRR [6, 16].

To determine the feasibility of using an automated classifier for IRR, we selected three themes and their associated words from our team's discussion: distance-learning, teaching-learning, and district-policies. We provided a classifier (nCoder, [16]) with the codes and keywords, "trained" the classifier through human coding of a training set (80 lines/code), had the classifier automatically code the data corpus, and compared codes from a test set to establish IRR. In practice, the process of establishing IRR between researchers, or researchers and computational approaches is iterative. The small size of the training set is only to explore the potential of using the automated classifier for dependability.

The IRR between the classifier and the researcher varied, $\kappa = (.35–.72)$, $\rho > .05$. For codes with low reliability, we found that clusters with fewer keywords (e.g., "distance learning" in deductive code) were harder to establish high reliability for, $\kappa = .35$; $\rho(.65) = 1$. Common phrases (e.g., "project", "expectations") could appear in multiple contexts, and thus there were high rates of false positives for codes that contained these words as identifiers (i.e., precision $< .60$).

Strengthening Dependability. We found that establishing high inter-rater reliability between the automated classifier and researchers was challenging when the training keywords from the algorithms were not unique to the codes. Prior work recommends that in the event of low agreement, the researchers can include more regular expressions for training, while updating the code definitions [6]. Developing and refining keywords is a nuanced process that is rarely documented in study write-ups. To strengthen notions of dependability in leveraging automated classifiers, instead of reporting only the final inter-rater statistics, researchers should document the changes to the code definitions as they work on establishing substantial agreement in automated analyses.

7 Triangulating Findings with Visual Analytics

Visual analytics techniques can serve as another way to show triangulation of data and interpretive process.

Our next insight toward establishing trustworthiness concerns use of visual analytics techniques, such as Epistemic Network Analysis (ENA) [24], to triangulate findings

from algorithmic approaches. To illustrate this workflow, we took "themes" that arose after our research team analyzed the word clusters from the two NLP algorithms. For example, one theme was teaching-and-learning and another was distance-learning. We then performed an ENA analysis to examine whether these themes were present, for a given interview excerpt.

For each sentence in the text, ENA counts the occurrences of the themes' related keywords in a window of conversational turns. To bound excerpts, we chose a moving window of 4 sentences. If two or more themes (via their keywords) are present in a text excerpt, they are co-present and reflected as connected in the network graph (see Fig. 2). ENA then normalizes the networks and projects them onto a lower dimensional space. Each node represents a theme (e.g., distance learning); thicker and darker lines represent higher frequencies of occurrences; and subtracting the networks results in comparison graphs that illustrate the differences between networks.

The ENA visualizations allowed us to confirm that different algorithmic approaches to parsing and categorizing words, led to different types of connections between themes. For example, the differences between the two approaches (Fig. 2, A and B) appeared in the positions of specific educator roles (i.e., nodes). This suggests that the conceptualization of each role by the three themes (teaching and learning, distance learning, and district intervention) appeared to differ between the two approaches.

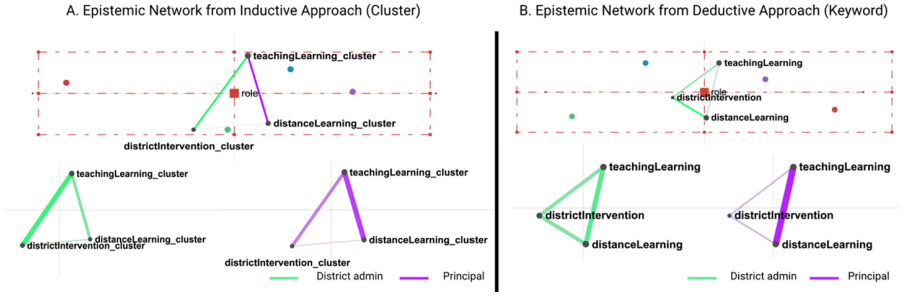

Fig. 2. Epistemic Networks for each Analysis Approach. *Notes*. Top: differences between principal-district networks; bottom: district (green) and principal (purple). (Color figure online)

Notably, ENA allowed for comparisons of different connections of themes based on whether a district administrator or a school principal was voicing these ideas. For example, Fig. 2.A shows ENA graphs based on the inductive, algorithmic model we used. On the lower left (green graph) we see that district administrators talked about teaching and learning and district interventions, more often together. On the lower right (purple graph), we observe that school principals often mentioned teaching and learning and distance learning, more often together. Could these utterances occur because of the roles that these staff play in a school system? These questions helped our team to revisit our qualitative data corpus, from different lenses, and combined with previous approaches offer multiple layers from which to develop insights.

Strengthening Credibility. Using other analysis strategies and visual analytics such as ENA, can strengthen arguments for credibility in qualitative research. Specifically, algorithmic and ENA approaches can be seen as other forms of triangulation – of looking at data from different viewpoints – and develop more rigorous claims and interpretations from data to strengthen the overall analysis.

8 Conclusions

In this paper, we examine how combining output from NLP algorithms, or what one research collaborator playfully described as "word vomit from the machine", with human interpretations can offer insights for establishing trustworthiness for the research process. We document our algorithmic processes to provide transparency to the analysis pipelines. Our experiences shed light on how different approaches influenced disparate research directions and analyses. We also highlight strategies to triangulate findings with visual analytics tools and establish inter-rater reliability (IRR) with an automated classifier. The iterativity in triangulating findings and establishing IRR implies an interdependence between the human and the algorithms that is crucial to the final interpretations.

Table 3. Algorithmic affordances to establish trustworthiness.

Facets	Definition	Use of computation	Our study
Credibility	Are the presented data plausible?	Algorithmic transparency triangulation Constant comparisons	2 algorithmic approaches Visual analytics
Dependability	Are the processes consistent & reliable?	Inter-rater reliability Pre-register analyses	Automated classifier
Confirmability	Are the findings traceable to original data?	Code Mapping	Human-algorithm interpretation

We return to our original premise: How can we establish trustworthiness in qualitative research when adopting human-computer analytic approaches? Table 3 summarizes several strategies for reporting and interpreting human-computer qualitative work. While process of triangulation, constant comparisons, establishing IRR, code mapping, and pre-registering analyses are geared towards analysts conducting the research, we also discuss how algorithmic transparency can help readers confirm the work's validity and bring in researchers who understand trustworthiness and the theoretical underpinnings of qualitative work, but may not be familiar with algorithmic workflows.

The current work is limited to our data contexts and participant insights, and we only explored two ways to code data in our illustrative example. Still, our processes of data collection, analyses, and human interpretation illuminate how different algorithmic approaches can aid interpretations of large qualitative data corpus. Directions for future research include efforts to build synergistic interfaces for researchers to conduct data exploration, interpretation, and triangulation with different computational techniques.

References

1. Anfara Jr, V.A., Brown, K.M., Mangione, T.L.: Qualitative analysis on stage: Making the research process more public. Educ. Res. **31**(7), 28–38 (2002)
2. Bakharia, A.: On the equivalence of inductive content analysis and topic modeling. In: Eagan, B., Misfeldt, M., Siebert-Evenstone, A. (eds.) ICQE 2019. CCIS, vol. 1112, pp. 291–298. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33232-7_25
3. Bakharia, A., Corrin, L.: Using recent advances in contextual word embeddings to improve the quantitative ethnography workflow. In: Eagan, B., Misfeldt, M., Siebert-Evenstone, A. (eds.) ICQE 2019. CCIS, vol. 1112, pp. 299–306. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33232-7_26
4. Boyd, R.L.: Psychological text analysis in the digital humanities. In: Hai-Jew, S. (ed.) Analytics in Digital Humanities. Multimedia Systems and Applications, Springer, Cham (2019). https://doi.org/10.1007/978-3-319-54499-1_7
5. Bullough, R.V., Jr., Pinnegar, S.: Guidelines for quality in autobiographical forms of self-study research. Educ. Res. **30**(3), 13–21 (2001)
6. Cai, Z., Siebert-Evenstone, A., Eagan, B., Shaffer, D.W., Hu, X., Graesser, A.C.: nCoder+: a semantic tool for improving recall of nCoder coding. In: Eagan, B., Misfeldt, M., Siebert-Evenstone, A. (eds.) ICQE 2019. CCIS, vol. 1112, pp. 41–54. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33232-7_4
7. Charmaz, K., Denzin, N.K., Lincoln, Y.S.: Handbook of qualitative research. Sage Publications, Thousand Oaks (2000)
8. Chen, N.C., Drouhard, M., Kocielnik, R., Suh, J., Aragon, C.R.: Using machine learning to support qualitative coding in social science: shifting the focus to ambiguity. ACM Trans. Interact. Intell. Syst. (TiiS) **8**(2), 1–20 (2018)
9. DiMaggio, P., Nag, M., Blei, D.: Exploiting affinities between topic modeling and the sociological perspective on culture: application to newspaper coverage of US government arts funding. Poetics **41**(6), 570–606 (2013)
10. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. KDD **96**(34), 226–231 (1996)
11. Fekete, J. D., Dufournaud, N.: Compus: visualization and analysis of structured documents for understanding social life in the 16th century. In: Proceedings of the Fifth ACM Conference on Digital Libraries, pp. 47–55 (2000)
12. Goldberg, Y.: Neural network methods for natural language processing. Synth. Lect. Hum. Lang. Technol. **10**(1), 1–309 (2017)
13. Harris, Z.S.: Distributional structure. Word **10**(2–3), 146–162 (1954)
14. Li, H., Schnieders, J.Z.Y., Bobek, B.L.: Theme analyses for open-ended survey responses in education research on summer melt phenomenon. In: Eagan, B., Misfeldt, M., Siebert-Evenstone, A. (eds.) ICQE 2019. CCIS, vol. 1112, pp. 128–140. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33232-7_11
15. Lincoln, Y.S., Guba, E.G.: But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation. New Direct. Program Eval. **1986**(30), pp. 73–84 (1986)
16. Marquart, C.L., Swiecki, Z., Eagan, B., Shaffer, D.W.: ncodeR (Version 0.1.2) (2018). https://cran.r-project.org/web/packages/ncodeR/ncodeR.pdf
17. Marshall, C., Rossman, G.B.: Designing Qualitative Research. Sage Publications, Thousand Oaks (2014)
18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J. Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)

19. Muralidharan, A., Hearst, M. A.: Supporting exploratory text analysis in literature study. Liter. Linguist. Comput. **28**(2), 283–295 (2012)
20. Ophir, Y., Walter, D., Marchant, E.R.: A collaborative way of knowing: bridging computational communication research and grounded theory ethnography. J. Commun. **70**(3), 447–472 (2020)
21. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
22. Rahmah, N., Sitanggang, I.S.: Determination of optimal epsilon (eps) value on DBscan algorithm to clustering data on peatland hotspots in sumatra. In IOP Conference Series: Earth and Environmental Science, vol. 31, no. 1, p. 012012. IOP Publishing (2016)
23. Robinson, R.L., Navea, R., Ickes, W.: Predicting final course performance from students' written self-introductions: a LIWC analysis. J. Lang. Soc. Psychol. **32**(4), 469–479 (2013)
24. Shaffer, D. W. Quantitative ethnography (2017). Lulu.com
25. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: LIWC and computerized text analysis methods. J. Lang. Soc. Psychol. **29**(1), 24–54 (2010)
26. Verma, M., Srivastava, M., Chack, N., Diswar, A.K., Gupta, N.: A comparative study of various clustering algorithms in data mining. Int. J. Eng. Res. Appl. (IJERA) **2**(3), 1379–1384 (2012)