The Computational Thematic Analysis Toolkit

ROBERT P. GAUTHIER, University of Waterloo, Canada
JAMES R. WALLACE, University of Waterloo, Canada

As online communities have grown, Computational Social Science has rapidly developed new techniques to study them. However, these techniques require researchers to become experts in a wide variety of tools in addition to qualitative and computational research methods. Studying online communities also requires researchers to constantly navigate highly contextual ethical and transparency considerations when engaging with data, such as respecting their members' privacy when discussing sensitive or stigmatized topics. To overcome these challenges, we developed the *Computational Thematic Analysis Toolkit*, a modular software package that supports analysis of online communities by combining aspects of reflexive thematic analysis with computational techniques. Our toolkit demonstrates how common analysis tasks like data collection, cleaning and filtering, modelling and sampling, and coding can be implemented within a single visual interface, and how that interface can encourage researchers to manage ethical and transparency considerations throughout their research process.

CCS Concepts: • **Computing methodologies**; • **Human-centered computing** → **Empirical studies in HCI**; **Collaborative and social computing theory, concepts and paradigms**;

Additional Key Words and Phrases: methodology; thematic analysis; computational methods; design

ACM Reference Format:
Robert P. Gauthier and James R. Wallace. 2022. The Computational Thematic Analysis Toolkit. *Proc. ACM Hum.-Comput. Interact.* 6, GROUP, Article 25 (January 2022), 15 pages. https://doi.org/10.1145/3492844

1 Introduction

Many researchers seek to develop a qualitative understanding of the activities taking place online, and have explored use of computational methods to help them overcome the volume, velocity, and variety of data found in those communities. In building on the strengths of qualitative and computational methods, researchers hope to enable a qualitative understanding of activities at scale that is not possible with traditional methods through a synthesis of techniques. For instance, the literature has explored augmenting qualitative techniques with natural language processing [23] of text, use of computational techniques like topic modelling (e.g. [6, 16, 17, 20, 39, 44]) to support data sampling [29], and development of classifiers to support humans in coding [13].

However, substantial barriers exist to this synthesis of methods in practice. For instance, researchers must be experts in qualitative research methodologies, like Grounded Theory [56] or Thematic Analysis [11], as well as computational techniques, like topic modelling [9], all of which require extensive training. Further, myriad ethical considerations need to be made when working with online community data, like whether the communities being studied have an expectation of privacy, the community's and their platform's terms of use, and balancing the need for transparent research

Authors' addresses: Robert P. Gauthier, University of Waterloo, Waterloo, Ontario, Canada, rpgauthier@uwaterloo.ca; James R. Wallace, University of Waterloo, Waterloo, Ontario, Canada, james.wallace@uwaterloo.ca.

2573-0142/2022/1-ART25 $15.00
https://doi.org/10.1145/3492844

practices with potential harm to the communities being studied [12, 26, 36]. Finally, researchers must navigate all of the these challenges across a number of separate tools; current software was not designed to support computational and qualitative methods in a single interface. Indeed, they must perform the substantial, iterative, and often tedious work associated with both qualitative research and computational methods; like data familiarization, cleaning, modelling, development of themes, and revisiting and iterating on those tasks when results are deemed unsatisfactory.

To overcome these barriers, we developed the *Computational Thematic Analysis Toolkit* which brings together elements of reflexive thematic analysis and computational methods under one cohesive, visual interface that is accessible to non-programmers. Our toolkit supports common analysis tasks like data collection from online communities (e.g., [7]), cleaning and filtering, modelling and sampling, and coding (e.g., [11]). In presenting our toolkit, we first define a conceptual workflow based on the many similarities between qualitative analysis and computational methods. We then present the toolkit's implementation for each analysis task. We also pay particular attention to guidance from the HCI literature on the integration of computational and qualitative research methods (e.g., [6, 13, 32, 42]), and how this guidance was implemented in our design.

2 Related Work

The Computational Social Science community has begun to explore the substantial similarities between qualitative analysis and computational methods. For instance, Muller et al. [42] notes 'convergences' between the two approaches to research; both are grounded in data, involve the creation of codes or features, are highly iterative, and that results are ultimately interpretive in nature. In discussing these similarities, they raise the question of whether hybrid methods might provide complementary benefits; that is "What if we could enjoy the virtues of both ways of inquiring?" [42, p.3].

Yet, despite their similarities, the tools used to support these different methods remain separate, and play distinct roles in their respective analyses. Qualitative researchers use a variety of tools [32] including office software, such as Excel [14] and Google Docs [33], as well as more specialized software like NVivo [34], Atlas.TI [27], or MAXQDA [55]. These tools typically have visual interfaces and are designed to play a supporting role rather then be the subject of analysis.

In contrast, computational tools typically require programming experience, like Mallet [38], Gensim [51], and TensorFlow [1]. They include both stand-alone software, as well as programming libraries that can be used alongside other tools. And while these libraries are often powerful and adaptable, there is some 'assembly required' before they can be used for analysis. Especially when compared to qualitative tools, computational tools play a central role in analysis; they are often integral to the analysis itself, rather than seen as supporting a human researcher.

Currently, this dichotomy in tools is a barrier to the integration of qualitative analysis and computational methods. They require different kinds of expertise. They are integrated with different tasks of their respective methods. It is also hard to transfer data between different tools, particularly given the highly iterative nature of both approaches to research. Thus, in this work, we consider how such techniques can be best integrated to support qualitative analysis of online communities. In particular, we focus on reflexive thematic analysis, as defined by Braun and Clarke [11].

2.1 Thematic Analysis and Computational Sampling

One of the most significant challenges to performing thematic analysis of online communities is how their scale impacts sampling and the time intensity of analysis [10, 11]; each community potentially comprises hundreds of thousands of posts from tens of thousands of people over a period of years. One approach to dealing with this size is to sample. For instance, researchers frequently use random selection [5] or convenience sampling, such as a date-window [2, 19, 28, 52], to obtain a sample that is small enough for human analysis. However, by doing so researchers risk missing interesting parts of the data before they can familiarize themselves with the data, perform coding, or review themes.

As such, computational techniques can assist with sampling and analysis tasks. Topic modelling techniques like Latent Dirichlet Allocation (LDA) [9] have been identified by the research community as a potential avenue for engagement with such large data sets [3, 13, 24, 35, 42]. In particular, researchers have begun exploring how topic models can be used to provide a lens into the data, be used to identify latent themes [35, 48], and to produce useful samples from large scale data [6, 29, 42]. If chosen wisely, computational techniques can assist with qualitative sampling strategies, such as purposeful sampling and theoretical sampling [37], and can help researchers develop an understanding of complex issues.

However, researchers also do not want computational methods to simply automate their interaction with the data; they want to maintain autonomy, intimacy, and ownership of their qualitative analysis [32]. In response to these needs, the research community has explored how humans might guide topic modelling techniques like LDA [21, 22]. Thus, there is a need for tools that support computational approaches to thematic analysis, but maintain the human researcher's role as the driver of analysis. In this work, we develop a conceptual framework for how qualitative and computational methods can be bridged in the context of reflexive thematic analysis, and create a toolkit that implements that framework.

2.2 Ethical and Transparency Considerations

Alongside the methodological and technical progress made by the research community, there has been a growing recognition that computational social science researchers need to actively consider the impacts of their processes, artifacts, and results [15]. That is, "ethics exist within a social context" [63, p. 77] and are intertwined with social considerations about how our research is performed and how it is used. In particular, research on digital communities can not be done in a one-size-fits-all manner [26]. While once such data might have been considered 'public' and safe, the research community now acknowledges that its use can put community members at risk [12, 36]. As such, there is a need for researchers to responsibly handle such research data, and for tools that support them in doing so.

There is also a growing recognition of the importance of transparency in Computational Social Science research, as well as within the broader HCI community (e.g., [57, 61, 62]). Transparent reporting strengthens the rigour and trustworthiness of research [49, 58, 60]. It helps others understand, evaluate, and build upon published work. It can also help the researchers *themselves*: it can help them familiarize themselves with, interpret, and reflect on the myriad decisions they make when working with data. Decisions like what data were captured and not captured? What was filtered during cleaning? And which topic models were created and why?

Despite transparency's importance, there remains little agreement on how it should be supported in research tools. When navigating epistemological tensions between interpretivist social science and positivist computational science perspectives there is often an emphasis on tying transparency to *replication* despite it being an inappropriate for many qualitative research studies, which instead seek to establish *trustworthiness* [49]. Indeed, complete transparency may also be inappropriate in some contexts. For instance, research involving marginalized or stigmatized groups may necessitate reduced transparency through paraphrasing [12] or fabricating data [36]. Thus, in developing our toolkit, we sought to embody these values within its design, and to show how ethical and transparency considerations might be embedded into the tools themselves.

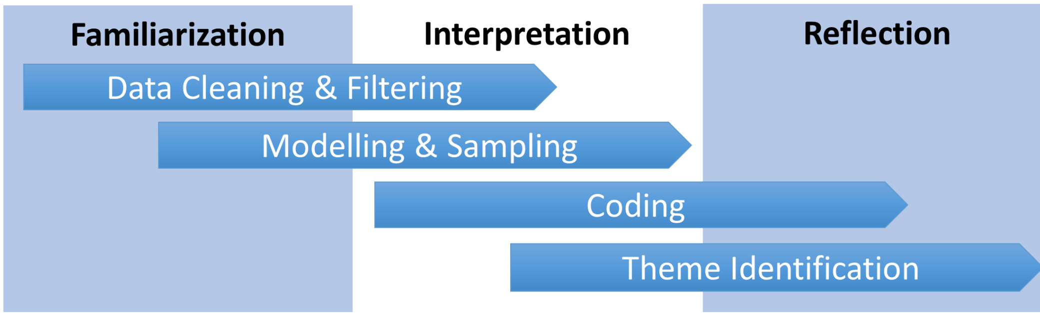

Fig. 1. Our Computational Thematic Analysis workflow. In moving from data collection to writing their final report, researchers progress through three conceptual stages of work: familiarization, interpretation, and reflection. To do so, they perform the practical tasks of data cleaning and filtering, modelling and sampling, coding, and theme identification. Like thematic analysis and computational methods, this workflow is highly iterative, and is not a linear process; researchers may shift between any conceptual stage or practical task.

3 Computational Thematic Analysis

Grounded in Braun and Clarke's [11] six-phase model of reflexive thematic analysis, we wanted to explore how computational techniques could play a larger role in supporting analysis by human researchers. To do so, we started by considering the commonalities, or 'convergences' [42], between thematic analysis and the computational methods used by the data science and machine learning communities. In particular, we considered similarities between how models and themes were developed, interpreted, and then used to identify patterns in the data [11, 50]. We then developed a hybrid workflow (Figure 1) that scaffolds the practical tasks associated with these analyses into three abstract, conceptual stages: familiarization, interpretation, and reflection.

We emphasize that while Figure 1 shows these conceptual stages and tasks in a linear manner, they are in practice highly iterative, and tasks frequently overlap. That is, as when performing thematic analysis or computational methods, researchers are likely to jump between conceptual stages and tasks as their analysis develops. For instance, a researcher might choose to code parts of their data during familiarization, and then sample before generating themes. In this way, the stages and tasks provide a guideline of how an analysis progresses, rather than a strict process that must be followed. We now overview each stage, and their relationships to the work performed in thematic analysis and computational methods.

3.0.1 Familiarization Regardless of whether researchers are approaching their analysis from the perspective of thematic analysis or computational methods, they must first familiarize themselves with the data; they need to confirm that they've collected appropriate data, spend time understanding it, and identify patterns [11]. In our workflow, familiarization encompasses tasks like data collection and inspection from thematic analysis, and pre-processing and exploration from computational methods. Computational techniques can augment these traditional thematic analysis tasks by automating aspects of data cleaning, providing an overview of collected data, and better enabling researchers to familiarize themselves with semantic, or word-level, patterns in the data.

3.0.2 Interpretation Interpretation is how researchers turn their familiarity with data into their own ideas. It involves tasks like sampling, generating initial codes, searching for themes, and theme identification from thematic analysis, and modelling and sampling from computational methods. The primary benefit of a hybrid model for these tasks is that computational techniques, like topic modelling, can be used iteratively by researchers to generate models, latent patterns within those models, and consider how the patterns suggest new assumptions and/or impact existing assumptions about their analysis [13, 24, 42], giving them access to a larger variety of patterns to investigate in later stages of thematic analysis. These techniques can also improve researchers' ability to describe the contextual nature of their assumptions and how their assumptions might influence their analysis.

While drawing a definitive line between familiarization and interpretation is difficult due to the highly iterative and overlapping nature of thematic analysis, it can be helpful to consider familiarization as being more focused on *semantic* patterns, such as reoccurring words or phrases, whereas interpretation then seeks to transition to more abstract understandings or *latent* patterns, such as reoccurring topics of discussion.

3.0.3 Reflection Reflection is where researchers consider whether their interpretations line up with the data, their experiences interacting with the data, and with their broader understanding of the domain of interest. It includes tasks like reflecting on codes, reviewing themes, and writing the final report. While we did not implement any computational techniques to support this aspect of thematic analysis, we posit that they may be used to suggest alternative interpretations of the data, for instance through summarization or counterfactuals.

4 The Computational Thematic Analysis Toolkit

To explore how our hybrid workflow and lessons learned from the literature might be embodied in a visual interface, we developed the *Computational Thematic Analysis Toolkit*. Imported data is visualized, cleaned, filtered, sampled in an interactive setting that enables rapid, iterative exploration and analysis. To support the inherent flexibility of our hybrid workflow, the toolkit's design is modular, where changes made at one stage of analysis are immediately reflected in the others, enabling researchers to shift between tasks as they progress between the familiarization, interpretation, and reflection stages of analysis.

Our toolkit supports each of the conceptual stages identified in our workflow (Figure 1), with each task assigned its own tab: Data Collection, Data Cleaning & Filtering, Modelling & Sampling, and Coding. The Data Collection tab enables researchers to download a data from various sources. The Data Cleaning & Filtering tab enables researchers to inspect, clean, and filter the data. The Modelling & Sampling tab enables researchers to locate data of interest for their thematic analysis. In particular, it involves applying machine learning approaches to the data to create and visualize samples of potential interest. The Coding tab provides a central location to develop, apply, and review codes by using data identified during both sampling and manual data inspection.

To maintain control of research data we implemented our toolkit as a standalone application instead of as a web application. We implemented our toolkit using Python 3 due to the wide availability and interoperability of packages for computational techniques, visualization, and user interfaces. Reddit data was obtained from pushshift.io's web API [7]. Individual data analysis components were implemented using pandas [47], spaCy [30], NLTK [8], Gensim [51] and biterm [59]. The toolkit's GUI was implemented using wxPython [18], with visualizations supported by Matplotlib [31], wordcloud [41], and mpl-chord-diagram [25]. The toolkit's full source code and installation instructions are available at https://osf.io/b72dm/.

4.1 Data Collection

To focus on the data, we moved from data collection being a code-oriented workflow to a visual interface that displays the data directly (Figure 2), similar to existing analysis software like *Tableau* where data is imported into a large collection that can then be visually modified or inspected. This

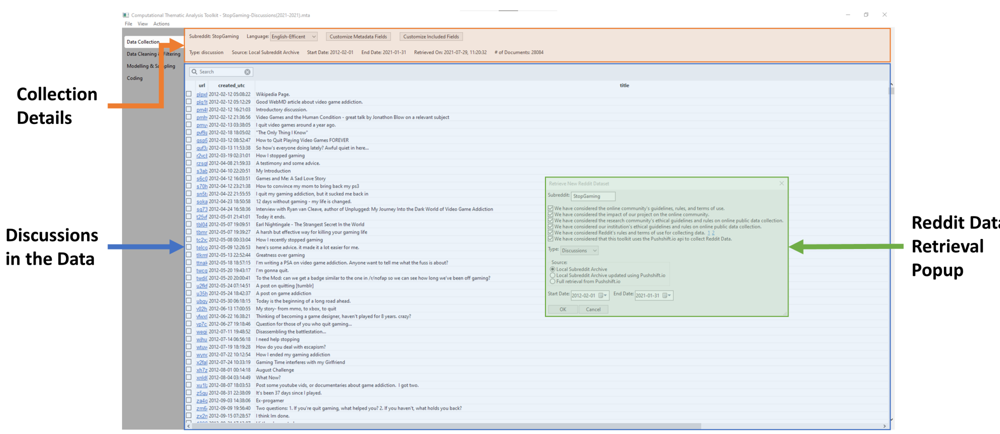

Fig. 2. Our toolkit's Data Collection tab provides a visual interface for imported discussion data. We automated the importation workflow, to enable researchers to pull data from sources like Reddit or Twitter without needing to write code while respecting each platform's terms of use, as well as the individual community's expectations for use of their data. Once imported, the data is available in for use in each of the other tabs.

approach also parallels computational tools like *Jupyter Notebook*, albeit with a less programming-centric metaphor for interaction. Our toolkit automates many aspects of connecting to data APIs for platforms such as Reddit and Twitter through visual dialogues, and enables researchers to directly import data from previous projects.

4.1.1 Importing Data A typical data import workflow includes four steps: (1) Import the submissions, (2) Import the comments that responded to those submissions, (3) group the submissions and comments using their `id` and `submission_id` fields, respectively, to form discussions and (4) select fields from those discussions for analysis. Each of these steps involves nuanced decisions, for instance, initial submissions and responses may fall on different sides of the defined time period for sampling. The workflow also needs to account for some fields being out-of-date when pulling from sources like pushshift.io, since they rely on snapshots taken shortly after a submission is created. For instance, fields like `author`, `created_utc`, `title`, `selftext`, and body are less likely to change after the archive was created, whereas fields like `score` (upvotes and downvotes) and `num_crossposts` tended not to be reliable in the archive.

To handle this complexity we developed an automated workflow for Reddit discussions, so that submissions and their responses are automatically grouped together when imported based on a given time period. We only include submissions made after the defined start time, and responses to those submissions that were made before the defined end time. We also selected a default subset of more stable fields to import, based on those we found useful during our thematic analyses: `URL`, `created_utc`, `title`, `text`, and body. However, acknowledging that thematic analysis by definition requires flexibility, the included fields are all modifiable. To be transparent about these concerns, the import dialogue includes warnings for each field that is unreliable when pulled from an archive.

4.1.2 Ethical Considerations and Retrieval Costs Ethical considerations are also deeply integrated with how data is first imported into the toolkit. We began by recognizing that there is no one-size-fits-all approach to handling these ethical considerations [26], and that our toolkit should encourage researchers to navigate their ethical choices rather than rigidly enforce a single set of decisions for all analyses. We then reviewed the literature and identified the following general opportunities for design: (1) communities' position on the use of their data by researchers, (2) ethical guidelines and rules of institutions, governments, and the research community [12, 15, 43], (3) the need to respect community actions such as moderation, (4) the need to respect individuals' actions such as deleting posts and/or accounts, and (5) the terms of use and capacity of the platforms we use to retrieve the data.

To encourage researchers to actively reflect on the five considerations above, the toolkit's data retrieval dialog implements friction in the form of confirmation check boxes. For example, in the case of Reddit, we implement multiple ethics confirmation check boxes (Figure 2), which provoke researchers to consider and reflect on how their analysis might impact different groups, including the community being studied, the research community, and Reddit.

We also implemented local caching of data to help reduce the amount of requests made to the platforms' APIs, and to improve the performance of computational techniques. However, even though this caching is technically permitted by both Pushshift and Reddit, it can be in tension with respecting the actions of both communities and individuals. That is, the cache might become out-of-sync, and contain posts that have been moderated by the community, or deleted by the person who made them. To help researchers identify and manage out-of-sync data, our toolkit links each discussion to its online source, providing a means of manually confirming which posts should be included in the analysis.

4.2 Data Cleaning & Filtering

We also developed a visual interface for cleaning and filtering tasks (Figure 3), and to enable researchers to become familiar with and interpret various patterns during their thematic analysis. In addition to these tasks being prerequisites for later computational techniques and machine learning tasks [50], tight integration of cleaning and filtering into a thematic analysis workflow facilitates the discovery of semantic patterns in the data. Our interface also provides opportunities to support transparency and interpretability of the computational aspects of the thematic analysis process.

Our toolkit includes a default data cleaning workflow, based on off-the-shelf tools available in Python. Imported data is tokenized using NLTK [31]. Tokens are then converted into both a stemmed form, using NLTK's SnowballStemmer [31], and a lemmatized form as well as tagged for parts-of-speech, and stop words status, using spaCy's pre-trained model [30]. The toolkit then computes descriptive data for each token, and stores them into pandas dataframes [47], including number of occurrences, number of discussions containing each token, and tf-idf. Should a researcher choose, each of these steps is modifiable.

4.2.1 Token-based Analysis To support data cleaning and familiarization, our toolkit's Data Cleaning & Filtering tab (Figure 3) provides list views for each token's part-of-speech, frequency, and tf-idf score. These views show how NLP is being used to interpret the data, and enable researchers to question underlying assumptions and fine-tune default settings before moving on to modelling and sampling tasks. These views are automatically updated when data is loaded and as filtering rules are adjusted. Search boxes enable researchers to quickly search the list for known words.

In developing and using the toolkit, we also found that visual inspection and review of these steps can inform qualitative analysis of data. While originally this interface was focused on helping us conduct modelling, as our toolkit evolved we found that the filtering and cleaning tasks enabled us to identify semantic patterns in the data. That is, looking at these lists also helped us familiarize ourselves with the data for our thematic analysis. For instance, these lists can be helpful in identifying domain-specific nouns that can be used as features in any models used to inspect the data, or

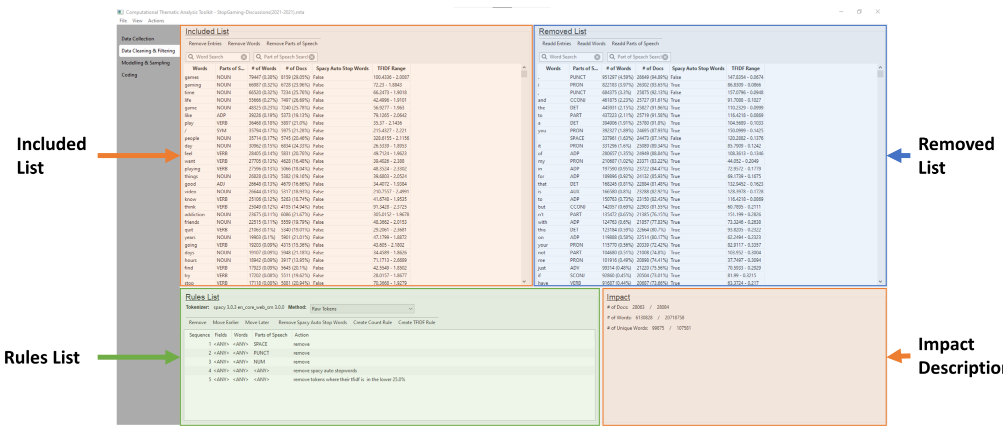

Fig. 3. Our toolkit's Data Cleaning & Filtering tab presents a summary of tokens present in the imported data. Researchers can use two list views to inspect and reflect on words included and removed by the data cleaning process. They can also manually fine-tune filters. We also developed a rules list that shows the cleaning and filtering steps applied, including their order, to assist with interpretability and transparency of the research process. To further transparency, our toolkit shows the impact filtering has on the data in terms of word counts and discussion counts to highlight the amount of data removed.

for coding and theme development. We have found that the list of *removed* words is equally useful in reflecting on and questioning underlying assumptions of our analysis.

4.2.2 Supporting Explainability & Transparency When using computational methods like *Juypter Notebooks*, analysts are initially aware of why a word was included or removed from their model(s), since each step is performed manually. However, much of this work becomes hidden when it is automated, making it difficult to determine which tokens had been included or removed, and in which order operations were applied. To make these decisions more transparent we implemented a rules list (Figure 3) that visualizes each operation and the order in which they are applied. To be transparent about the impact of filtering on the data, we also provide summaries of how many discussions, words, and unique words are present after all of the filtering rules have been applied.

4.3 Modelling & Sampling

Once the researcher has familiarized themselves with semantic patterns in the data, our Modelling & Sampling tab (Figure 4) provides functionality to help them identify and interpret latent patterns. It enables researchers to quickly generate, inspect, and interpret samples from computational models of their data. Researchers can generate unique models in separate tabs across the top of the display using random sampling, or purposive computational techniques like LDA and biterm, for long and short text, respectively. As models are generated, they are also visualized on the right-hand side of the screen, with samples shown in the bottom left, where they may be inspected by the researcher.

4.3.1 Visualizing Models While much of the topic modelling literature has focused on statistical evaluation of models [3, 4, 9, 40, 46, 53], when we tried to implement such measures into our qualitative research process we found that they were very abstract, difficult to explain, and difficult to apply. Additionally, we did not find a clear consensus on what metrics to use to optimize a model since each had their own strengths and weaknesses. We instead decided to focus on visualizing the model, to enable researchers to more rapidly understand, refine, and enhance their topics for qualitative interpretation [21, 22]. By evaluating the models visually we found that we could

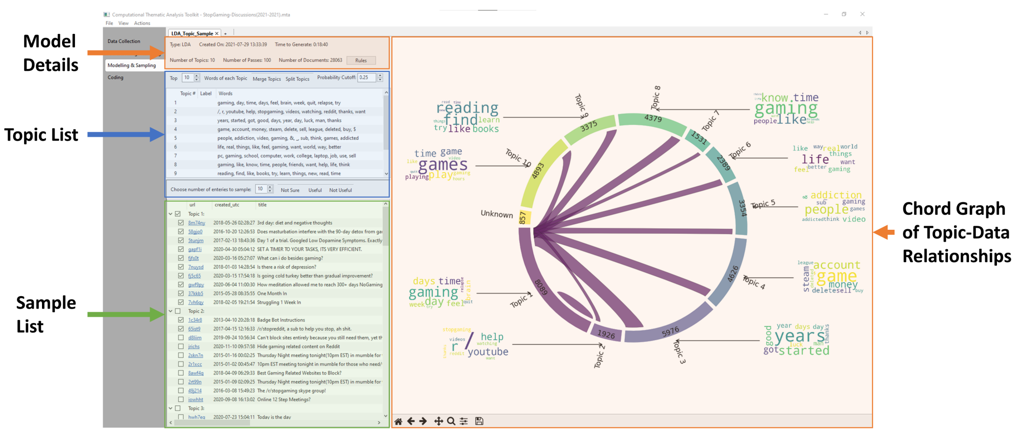

Fig. 4. The Modelling & Sampling Tab enables researchers to create topic models of the data set and visualize them using our chord graph. The graph shows topics, their respective keywords, and the portion of discussions shared with each other topic. Researchers may also label and merge topics created by the model before selecting discussions from the generated sample for coding and theme development.

generate models that are 'good enough' to hone in on interesting data, rather than getting caught up in trying to create an optimal model for some quantitative metric that didn't ultimately improve the thematic analysis.

In particular, our visualizations focused on both topic-discussion and topic-word relationships. Initially, we tried including pyLDAvis [54], a popular visualization for LDA models that shows whether topics overlap and how different words contribute to each topic. However, we found that it didn't help us explore how words interrelated between topics or identify when discussions might contain to multiple topics — particularly important questions when performing a thematic analysis. To better support these aspects of thematic analysis, we decided to implement our own visualization.

We created a chord graph (Figure 4) where each topic is represented by an arc, with its length corresponding to the portion of discussions it represents. Each arc is then labelled with a name and a word cloud containing the most relevant keywords for each topic. The visualization is also interactive. Researchers can click on any topic arc to view chords corresponding to discussions in common with other topics. Researchers can then mouse over other topics to make comparisons between more than one topic at once. Topics can also be named and merged by the researcher, as they familiarize themselves with the model and data. By default, topic names are blank so that they may be created by the researchers rather than the model [32].

4.3.2 Identifying Samples To facilitate sampling, our toolkit uses the generated models to identify sets of discussions to be used as samples for thematic analysis using the following steps: (1) The topic model is used to calculate the probability that each topic is present within each discussion. (2) A sorted list is created for each topic, with the most representative discussions at the top of the list. (3) Any discussion with a probability lower than the cutoff is dropped from each topic's list. (4) An 'unknown' topic is formed using any discussions whose topic probabilities are all below the cutoff. (5) For the 'unknown' topic, discussions are sorted based on *lowest* maximum probability to seek out discussions that the model is least confident in. (6) The top discussions from each topic are then included in the samples.

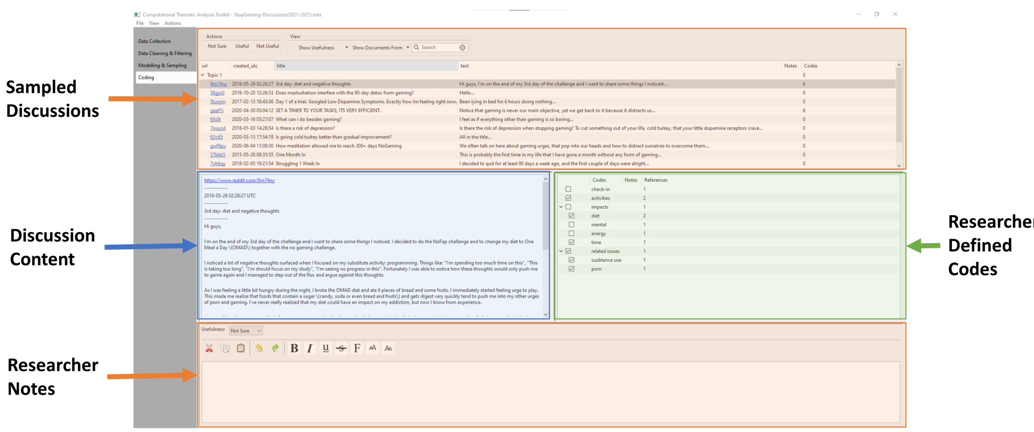

Fig. 5. Our toolkit's Coding tab displays a list of discussions along with, upon selecting a discussion, showing the discussion's collected fields, a list of codes researchers can assign to the discussion, and a text editor where researchers can make and review any notes about the discussion. Adding codes to a discussion can be done by either by checking off existing codes used in other discussions or creating new codes.

These samples are visualized in a list that shows each of the fields of the discussions on the lower left hand side of each Modelling & Sampling tab (Figure 4, Sample List). Researchers can interactively control how many discussions are sampled for each topic, and flag documents that they find useful or not-useful.

4.4 Coding

The Coding tab (Figure 5) facilitates researchers' interpretation of data through reviewing discussions, taking notes and applying codes. Rather than this tab using additional computational techniques, such as AI coders, we focused on providing researchers access to discussions identified manually, in the data collection tab, as well as from model-derived samples. This approach aligns with the goal of 'supporting serendipity' to maintain researchers' agency in the thematic analysis process [32]. Our toolkit supports three coding activities: reviewing discussions, applying codes, and writing notes.

Discussions are gathered when a researcher checks off desired discussions on the's activity during Data Collection and Modelling & Sampling tabs. Any discussion that a researcher has checked off is made available for coding in a list at the top the screen. Researchers can also search and toggle which discussions are displayed, based on those flagged as useful, unsure, and not useful. When a discussion is selected, the lower portion of the screen is used to perform coding: the full discussion content is displayed, codes can be applied, and researcher notes can be attached.

This coding tab is designed to mirror existing qualitative research software like NVivo [34] and Atlas.TI [27]. As researchers review each discussion, they can apply self-defined codes to them. Researcher-defined codes are displayed in a pane to the right, along with the number of discussions that have been tagged with that code. To support note taking, the coding interface provides a rich text interface with common word processing functions at the bottom of the screen.

5 Discussion

Our toolkit integrates support for qualitative and computational research methods within a single software interface. We built on the 'convergences' between qualitative and computational methods

[42], and from those convergences derived a computational thematic analysis workflow and a visual interface that enables non-technical researchers to engage in its methods. As the first implementation of these ideas, we hope that our toolkit can spark discussion in the field, and in particular, discussion around two aspects of its design: the interpretive role of computational support for qualitative analysis, and how to provide built-in support ethical and transparent research.

When integrating qualitative and computational techniques, we embraced Baumer et al.'s [6] view that models provide "scaffolding for human interpretation" and the importance of one or more *good enough* models which provide new perspectives of the data, rather than a singular, "best" model as an objective truth. Similarly, we emphasized a visual approach to model evaluation, eschewing quantifications like coherence [40, 46, 53] or Jaccard's distance [3]. These visualizations intentionally show data that was *not* associated with a topic to help researchers keep in mind that models are imperfect, and should be questioned and interpreted.

We also considered calls for ethical design and use of computational methods from the research community, and notably Jiang et al. [32]. We acknowledge these risks, but also feel that computational techniques can help to identify different understandings of a phenomenon, often in ways that may not be immediately accessible to a human researcher alone. To mitigate these concerns, we emphasized transparency — we designed our toolkit to support transparent reporting within each conceptual stage and practical analytic task to strengthen the rigour and trustworthiness of the research [58, 60].

6 Limitations

We designed our toolkit around *reflexive* thematic analysis [11], however other variations of thematic analysis (or more generally, other qualitative research methods) may not be as well supported by it. We believe that much of our toolkit's core functionality can be useful for different methods, and we designed it to be modular and flexible. But, if one wanted to, for example, begin with a qualitative analysis (i.e., coding) and then use those codes to create a supervised model for topic classification, then the current toolkit would not provide adequate support.

Further, additional research and development will be required to show the toolkit's efficacy in practice. This research will need to focus on both specific features and the general utility of the toolkit. For instance, novel features like our chord graph visualization require further study. Future work could perform comparative evaluations of our visualization and pyLDAvis [54] to elucidate their respective strengths and weaknesses. We also anticipate performing hands-on field studies with qualitative researchers to better understand our toolkit's use in practice.

Finally, and perhaps most significantly, we acknowledge that the ethical and transparent treatment of research data, and indeed the methods themselves, are a rapidly evolving research area. Our implementation is itself therefore a limitation of this research. In developing our toolkit, we sought to embody the values of transparency and ethical conduct into the software's design. Sometimes that meant adding friction, like when prompting researchers to pause and think about the potential consequences of their actions. We did this intentionally to align with what we felt were emerging best practices; but as this area of research is rapidly evolving and future work is likely to develop new best practices.

We also want to explicitly resist any notion of situating our toolkit as 'ethical', or the false dichotomy of other tools as 'unethical'. We instead acknowledge there are ethical risks involved when conducting research into online communities. Due to these risks, we aimed to make our tool provoke ethical considerations which enable researchers take control of their own ethics in research process. However, our implementation is certainly not perfect, nor complete. We only hope that it can serve as a working example and a point of critique for how ethics and transparency can be more deeply integrated into the software used to perform Computational Social Science.

7 Conclusion

In this work we set out to explore how computational techniques can augment human qualitative analysis, with a focus on reflexive thematic analysis. To do so, we developed the Computational Thematic Analysis toolkit to explore different types of support, and how those supports might be extended to researchers without expertise in computational techniques. We also embedded ethical considerations of data analysis into the toolkit, with the goal of encouraging researchers to consider the potential impacts of their research on the communities they study, and to support research transparency through records of both the decisions made for both qualitative and quantitative methods.

Finally, this research was exploratory by nature, intended to explore some of the 'big effects' [45] in designing for a hybrid qualitative and computational analysis. Our toolkit implements software features to support several tasks as defined in our Computational Thematic Analysis workflow, but unexplored supports, particularly in later tasks like 'theme identification', may also prove useful. For instance, future work might explore how counter-factuals can prompt a qualitative researcher to consider new perspectives on the themes that they had developed, or how summarization techniques might be used to help define and name themes later on in an analysis. We hope that by sharing our code under an open source license that this work can serve as a platform for the exploration of these ideas in future work.

8 Acknowledgements

We thank the University of Waterloo Games Institute and its members for their constructive feedback. This work was made possible by NSERC Discovery Grant 2015-06585 and the Ontario Graduate Scholarship.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, and et al. 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. http://arxiv.org/abs/1603.04467 arXiv: 1603.04467.

[2] Salim Ahmed, Rebecca J. Haines-Saah, Arfan R. Afzal, Helen Tam-Tham, Mohammad Al Mamun, Brenda R. Hemmelgarn, and Tanvir C. Turin. 2017. User-driven conversations about dialysis through Facebook: A qualitative thematic analysis. *Nephrology* 22, 4 (2017), 301–307. https://doi.org/10.1111/nep.12780 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/nep.12780.

[3] Tawfiq Ammari, Sarita Schoenebeck, and Daniel M. Romero. 2018. Pseudonymous Parents: Comparing Parenting Roles and Identities on the Mommit and Daddit Subreddits. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 489:1–489:13. https://doi.org/10.1145/3173574.3174063

[4] R. Arun, V. Suresh, C. E. Veni Madhavan, and M. N. Narasimha Murthy. 2010. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In *Advances in Knowledge Discovery and Data Mining (Lecture Notes in Computer Science)*, Mohammed J. Zaki, Jeffrey Xu Yu, B. Ravindran, and Vikram Pudi (Eds.). Springer, USA, 391–402. https://doi.org/10.1007/978-3-642-13657-3_43

[5] Angelica Attar and Neil S. Coulson. 2012. A thematic analysis of patient communication in Parkinson's disease online support group discussion forums. *Computers in Human Behavior* 28, 2 (Mar 2012), 500–506. https://doi.org/10.1016/j.chb.2011.10.022

[6] Eric P. S. Baumer, David Mimno, Shion Guha, Emily Quan, and Geri K. Gay. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology* 68, 6 (2017), 1397–1410. https://doi.org/10.1002/asi.23786 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23786

[7] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media* 14 (May 2020), 830–839.

[8] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* O'Reilly Media, Inc., Sebastopol, CA.

[9] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[10] Richard E. Boyatzis. 1998. *Transforming qualitative information: Thematic analysis and code development.* Sage Publications, Inc, Thousand Oaks, CA, US.

[11] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[12] Amy Bruckman. 2002. Studying the amateur artist: A perspective on disguising data collected in human subjects research on the Internet. *Ethics and Information Technology* 4, 3 (2002), 217–231. https://doi.org/10.1023/A:1021316409277

[13] Nan-Chen Chen, Margaret Drouhard, Rafal Kocielnik, Jina Suh, and Cecilia R. Aragon. 2018. Using Machine Learning to Support Qualitative Coding in Social Science: Shifting the Focus to Ambiguity. *ACM Trans. Interact. Intell. Syst.* 8, 2 (Jun 2018), 9:1–9:20. https://doi.org/10.1145/3185515

[14] Microsoft Corporation. 2021. Microsoft Excel. https://office.microsoft.com/excel

[15] Melissa Densmore, Casey Fiesler, Cosmin Munteanu, Michael Muller, Janet C. Read, Katie Shilton, and Özge Subaşı. 2020. Research Ethics Roundtable. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing (CSCW '20 Companion)*. Association for Computing Machinery, New York, New York, USA, 195–198. https://doi.org/10.1145/3406865.3419015

[16] Paul DiMaggio, Manish Nag, and David Blei. 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics* 41, 6 (2013), 570 – 606. https://doi.org/10.1016/j.poetic.2013.08.004 Topic Models and the Cultural Sciences.

[17] Karthik Dinakar, Birago Jones, Henry Lieberman, Rosalind Picard, Carolyn Rose, Matthew Thoman, and Roi Reichart. 2012. You too?! mixed-initiative lda story matching to help teens in distress. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM 2012)*. Association for the Advancement of Artificial Intelligence, Palo Alto, California, 74–81.

[18] Robin Dunn. 2021. wxPython: Cross platform GUI toolkit for Python, "Phoenix" version. http://wxPython.org/

[19] Alexandra R. D'Agostino, Allison R. Optican, Shaina J. Sowles, Melissa J. Krauss, Kiriam Escobar Lee, and Patricia A. Cavazos-Rehg. 2017. Social networking online to recover from opioid use disorder: A study of community interactions. *Drug and Alcohol Dependence* 181 (Dec. 2017), 5–10. https://doi.org/10.1016/j.drugalcdep.2017.09.010

[20] M. Eickhoff and Runhild Wieneke. 2018. Understanding Topic Models in Context: A Mixed-Methods Approach to the Meaningful Analysis of Large Document Collections. In *Proceedings of the 51st Hawaii International Conference on System Sciences*. University of Hawaii at Manoa, Honolulu, Hawaii, 903–912.

[21] Mennatallah El-Assady, Rita Sevastjanova, Fabian Sperrle, Daniel Keim, and Christopher Collins. 2018. Progressive Learning of Topic Modeling Parameters: A Visual Analytics Framework. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (Jan 2018), 382–391. https://doi.org/10.1109/TVCG.2017.2745080

[22] Mennatallah El-Assady, Fabian Sperrle, Oliver Deussen, Daniel Keim, and Christopher Collins. 2018. Visual Analytics for Topic Model Optimization based on User-Steerable Speculative Execution. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2018), 13. vialab.science.uoit.ca/wp-content/papercite-data/pdf/ela2018b2.pdf

[23] James A. Evans and Pedro Aceves. 2016. Machine Translation: Mining Text for Social Theory. *Annual Review of Sociology* 42, 1 (2016), 21–50. https://doi.org/10.1146/annurev-soc-081715-074206

[24] Jeanine Evers. 2018. Current Issues in Qualitative Data Analysis Software (QDAS): A User and Developer Perspective. *The Qualitative Report* 23, 13 (March 2018), 61–73. https://nsuworks.nova.edu/tqr/vol23/iss13/5

[25] Tanguy Fardet, cy1110, Francisco J. Martínez-Murcia, and Feng Wang. 2021. mpl-chord-diagram: Python module to plot chord diagrams with matplotlib. https://codeberg.org/tfardet/mpl_chord_diagram

[26] Casey Fiesler, Nathan Beard, and Brian C. Keegan. 2020. No Robots, Spiders, or Scrapers: Legal and Ethical Regulation of Data Collection Methods in Social Media Terms of Service. *Proceedings of the International AAAI Conference on Web and Social Media* 14 (May 2020), 187–196.

[27] Scientific Software Development GmbH. 2021. ATLAS.ti: The Qualitative Data Analysis & Research Software. https://atlasti.com/

[28] Rebecca J. Gooden and Helen R. Winefield. 2007. Breast and Prostate Cancer Online Discussion Boards: A Thematic Analysis of Gender Differences and Similarities. *Journal of Health Psychology* 12, 1 (Jan. 2007), 103–114. https://doi.org/10.1177/1359105307071744 Publisher: SAGE Publications Ltd.

[29] Orland Hoeber, Larena Hoeber, Ryan Snelgrove, and Laura Wood. 2017. Interactively Producing Purposive Samples for Qualitative Research using Exploratory Search.. In *SCST@ CHIIR*. Association for Computing Machinery, Oslo, Norway, 18–20.

[30] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. https://doi.org/10.5281/zenodo.1212303

[31] J. D. Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* 9, 3 (2007), 90–95. https://doi.org/10.1109/MCSE.2007.55

[32] Jialun Aaron Jiang, Kandrea Wade, Casey Fiesler, and Jed R. Brubaker. 2021. Supporting Serendipity: Opportunities and Challenges for Human-AI Collaboration in Qualitative Analysis. *Proceedings of the ACM on Human-Computer*

Interaction 5, CSCW1 (Apr 2021), 1–23. https://doi.org/10.1145/3449168 arXiv: 2102.03702.

[33] Google LLC. 2021. Google Docs. https://docs.google.com/

[34] QSR International Pty Ltd. 2021. NVivo qualitative data analysis software. https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

[35] Daniel Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler, H. Schmid-Petri, and S. Adam. 2018. Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Communication Methods and Measures* 12, 2-3 (2018), 93–118. https://doi.org/10.1080/19312458.2018.1430754 arXiv:https://doi.org/10.1080/19312458.2018.1430754

[36] Annette Markham. 2012. Fabrication as ethical practice: Qualitative inquiry in ambiguous internet contexts. *Information, Communication & Society* 15, 3 (2012), 334–353. https://doi.org/10.1080/1369118X.2011.641993

[37] Martin N. Marshall. 1996. Sampling for qualitative research. *Family Practice* 13, 6 (Jan 1996), 522–526. https://doi.org/10.1093/fampra/13.6.522

[38] Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu

[39] Yelena Mejova, Youcef Benkhedda, and Khairani Khairani. 2017. #Halal Culture on Instagram. *Frontiers in Digital Humanities* 4 (2017), 21. https://doi.org/10.3389/fdigh.2017.00021

[40] David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing Semantic Coherence in Topic Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Edinburgh, United Kingdom) *(EMNLP '11)*. Association for Computational Linguistics, USA, 262–272.

[41] Andreas Mueller. 2021. wordcloud: A little word cloud generator. https://github.com/amueller/word_cloud

[42] Michael Muller, Shion Guha, Eric P.S. Baumer, David Mimno, and N. Sadat Shami. 2016. Machine Learning and Grounded Theory Method: Convergence, Divergence, and Combination. In *Proceedings of the 19th International Conference on Supporting Group Work (GROUP '16)*. ACM, New York, NY, USA, 3–8. https://doi.org/10.1145/2957276.2957280

[43] Cosmin Munteanu, Amy Bruckman, Michael Muller, Christopher Frauenberger, Casey Fiesler, Robert E. Kraut, Katie Shilton, and Jenny Waycott. 2019. SIGCHI Research Ethics Town Hall. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. Association for Computing Machinery, New York, New York, USA, 1–6. https://doi.org/10.1145/3290607.3311742

[44] Laura K. Nelson. 2020. Computational Grounded Theory: A Methodological Framework. *Sociological Methods & Research* 49, 1 (2020), 3–42. https://doi.org/10.1177/0049124117729703 arXiv:https://doi.org/10.1177/0049124117729703

[45] Carman Neustaedter and Phoebe Sengers. 2012. Autobiographical design in HCI research: designing and learning through use-it-yourself. In *Proceedings of the Designing Interactive Systems Conference (DIS '12)*. Association for Computing Machinery, Newcastle Upon Tyne, United Kingdom, 514–523. https://doi.org/10.1145/2317956.2318034

[46] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic Evaluation of Topic Coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10)*. Association for Computational Linguistics, USA, 100–108. http://dl.acm.org/citation.cfm?id=1857999.1858011 event-place: Los Angeles, California.

[47] The pandas development team. 2020. pandas-dev/pandas: Pandas. https://doi.org/10.5281/zenodo.3509134

[48] Forough Poursabzi-Sangdeh, Jordan Boyd-Graber, Leah Findlater, and Kevin Seppi. 2016. Alto: Active learning with topic overviews for speeding label induction and document labeling. In *Proceedings of the 54th Annual Meeting of the (Volume 1: Long Papers)*, Vol. 1. Association for Computational Linguistics, Baltimore, MA, USA, 1158–1169. Issue 1.

[49] Michael G. Pratt, Sarah Kaplan, and Richard Whittington. 2020. Editorial Essay: The Tumult over Transparency: Decoupling Transparency from Replication in Establishing Trustworthy Qualitative Research. *Administrative Science Quarterly* 65, 1 (Mar 2020), 1–19. https://doi.org/10.1177/0001839219887663

[50] Sebastian Raschka. 2015. *Python Machine Learning*. Packt Publishing Ltd, USA. Google-Books-ID: GOVOCwAAQBAJ.

[51] Radim Rehurek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50.

[52] Shelly Rodgers and Qimei Chen. 2005. Internet Community Group Participation: Psychosocial Benefits for Women with Breast Cancer. *Journal of Computer-Mediated Communication* 10, JCMC1047 (Jul 2005), 1 pages. https://doi.org/10.1111/j.1083-6101.2005.tb00268.x

[53] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)*. ACM, New York, NY, USA, 399–408. https://doi.org/10.1145/2684822.2685324

[54] Carson Sievert and Kenneth Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*. Association for Computational Linguistics, Baltimore, Maryland, USA, 63–70.

[55] VERBI Software. 2021. MAXQDA. https://www.maxqda.com/qualitative-analysis-software

[56] Anselm Strauss and Juliet M Corbin. 1997. *Grounded theory in practice*. Sage, Thousand Oaks, CA.

[57] Poorna Talkad Sukumar, Ignacio Avellino, Christian Remy, Michael A. DeVito, Tawanna R. Dillahunt, Joanna McGrenere, and Max L. Wilson. 2020. Transparency in Qualitative Research: Increasing Fairness in the CHI Review Process. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3334480.3381066

[58] Poorna Talkad Sukumar, Ignacio Avellino, Christian Remy, Michael A. DeVito, Tawanna R. Dillahunt, Joanna McGrenere, and Max L. Wilson. 2020. Transparency in Qualitative Research: Increasing Fairness in the CHI Review Process. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3334480.3381066

[59] Markus Tretzmuller. 2021. biterm: Biterm Topic Model. https://github.com/markoarnauto/biterm

[60] Rivka Tuval-Mashiach. 20161027. Raising the curtain: The importance of transparency in qualitative research. *Qualitative Psychology* 4, 2 (20161027), 126. https://doi.org/10.1037/qup0000062

[61] Jan B. Vornhagen, April Tyack, and Elisa D. Mekler. 2020. Statistical Significance Testing at CHI PLAY: Challenges and Opportunities for More Transparency. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (Virtual Event, Canada) *(CHI PLAY '20)*. Association for Computing Machinery, New York, NY, USA, 4–18. https://doi.org/10.1145/3410404.3414229

[62] Chat Wacharamanotham, Lukas Eisenring, Steve Haroz, and Florian Echtler. 2020. *Transparency of CHI Research Artifacts: Results of a Self-Reported Survey*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376448

[63] Malcolm Williams. 2003. *Making Sense of Social Research*. SAGE Publications, Ltd, London, England, United Kingdom. https://doi.org/10.4135/9781849209434

Received July 2021; revised September 2021; accepted October 2021