Understanding Interaction with Machine Learning through a Thematic Analysis Coding Assistant: A User Study

FEDERICO MILANA, University College London, United Kingdom
ENRICO COSTANZA, University College London, United Kingdom
MIRCO MUSOLESI, University College London, United Kingdom
AMID AYOBI, University College London, United Kingdom

Interactive Machine Learning (ML) enables users, including non-experts in ML, to iteratively train and improve ML models. However, limited research has been reported on how non-experts interact with these systems. Focusing on thematic analysis as a practical application, we report on a user study where 20 participants interacted with TACA, a functioning Interactive ML tool. Thematic analysis involves individual interpretation of ambiguous data, hence it is suited for and can benefit from the iterative customization of models supported by Interactive ML. Through a combination of interaction logs and semi-structured interviews, our findings revealed that, by using TACA, participants critically reflected on their analysis, gained new thematic insights, and adapted their interpretative stance. We also document misconceptions of ML concepts, positivist views, and personal blame for poor model performance. We then discuss how applications could be designed to improve the understanding of Interactive ML concepts and foster reflexive work practices.

CCS Concepts: • **Human-centered computing → Interactive systems and tools**; • **Computing methodologies → Machine learning**.

Additional Key Words and Phrases: interactive machine learning, thematic analysis

ACM Reference Format:
Federico Milana, Enrico Costanza, Mirco Musolesi, and Amid Ayobi. 2025. Understanding Interaction with Machine Learning through a Thematic Analysis Coding Assistant: A User Study. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW197 (April 2025), 36 pages. https://doi.org/10.1145/3711095

1 INTRODUCTION

Machine learning (ML) has become ubiquitous in today's digital landscape, finding application in various domains and industries, and a growing number of users are interacting with systems that are driven by sophisticated algorithms. Currently, most of these applications are based on models trained on large data sets, with Large Language Models being the most recent examples. The dependence of model performance on the size of the training set is widely identified as one of the limitations of ML [44]. As a response, there is growing interest in achieving high performance by customizing models trained on smaller data sets [17, 40, 47, 70].

Interactive ML has been proposed as one approach to potentially achieve greater accuracy when training models on small data sets or on data that is ambiguous in nature [3]. Interactive ML involves the end-users in an iterative and incremental learning process and leverage human feedback to drive machine learning. Rapid iteration cycles of input, model updates and output allow the model to be fine-tuned incrementally by re-labeling misclassifications, labeling data points near decision boundaries or setting preferences and thresholds. This "human-in-the-loop" approach

Authors' addresses: Federico Milana, federico.milana.18@ucl.ac.uk, University College London, 66 - 72 Gower Street, London, United Kingdom, WC1E 6EA; Enrico Costanza, e.costanza@ucl.ac.uk, University College London, 66 - 72 Gower Street, London, United Kingdom; Mirco Musolesi, m.musolesi@ucl.ac.uk, University College London, London, United Kingdom; Amid Ayobi, amid.ayobi@ucl.ac.uk, University College London, 66 - 72 Gower Street, London, United Kingdom.

has already shown promising results in applications in health informatics [34] and environmental sciences [21], especially when only small data sets are available and in problems characterized by complex or rare events. Besides reducing the requirement for extensive data sets, another significant advantage is that model refinement can be driven by non-experts in ML [64]. While a wide range of users is now exposed to popular ML applications (e.g., ChatGPT), these do not generally allow users to intentionally train or refine models on their own data, and hence do not expose them to fundamental concepts.

The current body of research has focused on expert users or the usability of specific applications of Interactive ML, and lacks comprehensive insights into how non-expert end-users interact with these systems. We are therefore interested in exploring the understanding and behavior around Interactive ML models and applications, and focus on observing how non-experts in ML relate to the creation of models based on their own data.

The design of user studies around interactive AI systems has been recognized as a challenge [37], as it is important to support positive participants experiences in pursuing ecological validity. In this paper, we focus on qualitative data analysis (QDA) as a domain for a user study of Interactive ML. Prior work has highlighted the potential benefits of applying ML to QDA [16, 28]. Here, Interactive ML appears to be particularly suitable as it captures the individuality of the end-user by enabling the correction of ML models according to individual perspective and interpretation of the data. However, while QDA serves as a valuable application domain for our study, it should be emphasized that our main interest lies in users interaction with Interactive ML. It remains unclear how non-expert end-users understand and interact with these systems, including potential biases involved in the iterative process, so our work aims to address this particular research gap.

Our primary contribution is an improved understanding of how non-ML experts engage with and experience an Interactive ML system. To enable this, we designed and implemented a functional desktop application: the "Thematic Analysis Coding Assistant" (TACA). Thematic analysis is a QDA research method used to identify, analyze, and report patterns, or "themes", within data, involving an iterative process of reading the data, generating initial codes, grouping codes into themes, and reviewing themes. TACA allows users to import a qualitative data set and trains an ML classifier on an initial coding phase to suggest how the analysis can be extended by assigning the user-defined themes to sentences that were not previously coded. We designed the application to address reported limitations of Interactive ML UI by introducing novel approaches to data re-labeling and result visualization.

TACA enabled us to run a user study where 20 participants with previous experience in thematic analysis, but no experience nor training in ML, applied TACA to qualitative data from their own research (if available) or to a set of 21 newspaper restaurant reviews they were asked to analyze manually beforehand. They spent around 20 minutes interacting with the tool and took part in a following semi-structured interview after sharing the TACA interaction logs.

This work relates to the concept of Algorithmic Experience (AX), an analytical framework for understanding and improving user interactions with algorithms proposed by Alvarado and Waern [2]. Our study focuses on how users interact with and perceive ML processes, particularly in terms of "algorithmic awareness" and "algorithmic user control". Through our user study, we investigate participants' understanding of these processes and their engagement in activities that influence data re-classification.

Results show that TACA was effective in exposing our participants to Interactive ML and applying it on qualitative data sets. TACA's UI features promoted a thorough examination of the data, facilitated the evaluation of the model and the assignment of feedback during the Interactive ML cycle, but also led to some misconceptions regarding the functionality of the model. More significantly, our findings suggest that users with no experience in ML tend to perceive the model as an external, objective source of advice, and consequently hold themselves accountable when the model does not perform well. Based on this understanding, we discuss how technologies could be designed to support end-users in gaining an understanding of Interactive ML concepts and workflows and foster reflexive work practices beyond the scope of QDA.

2 RELATED WORK

Interactive ML aims to complement the computational power of ML algorithms with human intelligence by eliciting the user in rapid and fine-tuned iteration cycles of input, model updates and output. User input may vary between re-labeling misclassifications, providing and indicating representative samples and features, and setting preferences and thresholds [3, 22]. In contrast to conventional ML, the magnitude of each model update is typically small, focusing on a specific aspect of the model, meaning that a fast training algorithm is often preferred to strong induction [8, 25].

Despite requiring domain knowledge, model refinement can be driven by non-experts in ML, dismissing the traditional role of practitioners to collect, pre-process and transform the data, tune parameters of the learning algorithm, and assess the quality of the updated model [3]. Additionally, Interactive ML is less dependent on the size and quality of the training data set, potentially achieving a greater precision accuracy in less time and with less costs [8]. For these reasons, the "human-in-the-loop" approach has found success particularly in health informatics applications, such as bioimage analysis, genome annotation and protein folding, where human involvement is required to interpret complex or rare events correctly [11, 34, 63].

Compared to traditional ML-driven applications, Interactive ML is generally more aligned with the AX framework [2]. Traditional ML applications often operate as black boxes, where the underlying algorithms are hidden from users, providing little to no insight into how decisions are made. In contrast, AX emphasizes transparency and user control over algorithmic processes, and Interactive ML applications embody these principles by design. These are developed to provide users agency to participate in the refinement of the underlying algorithm with their own knowledge and preferences, and provide insights into how user behavior influences algorithmic outcomes. Our study specifically evaluates non-expert interactions within two of the five design areas of AX: "algorithmic awareness" and "algorithmic user control". "Algorithmic awareness" refers to the extent to which users understand the presence and role of algorithms in their interactions with the system. This includes recognizing when and how the ML model is influencing their experiences and outcomes. "Algorithmic user control", on the other hand, focuses on the tools to allow users to influence these algorithms, such as the feedback assignment phase of the Interactive ML cycle.

2.1 Interactive ML System Design

Addressing a lack of consolidated guidelines for Interactive ML system design is a review from Dudley *et al.*, who propose several solution principles following the four elements defined as: sample review, feedback assignment, model inspection, and task overview [22].

Not only is labeling data tedious and sometimes not considered worthwhile by the user, but it requires investing significant effort before noticeable change in the model [30, 51, 67]. Notably, there appears to be an opportunity in the evaluation of interaction techniques designed to enable the user to re-label multiple data points simultaneously. The presentation of representative and non-redundant samples could address both issues while allowing the user to assess the current state of the model more effectively.

In feedback assignment, the user manually selects features, re-assigns labels, or provides any other input designed to steer the model. Because constraints to the interactions with correction interfaces can easily translate to the degradation of the process, numerous studies have identified and explored novel interactions unrestricted to labeling instances, such as feature selection and weight adjustment [22, 48, 60]. However, these techniques pose significant interface design challenges to avoid overwhelming the user with too many, or too advanced, machine-centric metrics, whereas data labeling remains the most popular method for end-user input [3, 32].

Many possible causes of errors in ML fall under the categories of mislabeled data, feature deficiencies and insufficient data [4]. Several inspection techniques allow the user to detect failures and their sources differently, including presenting all of the unlabeled data points sorted by their predicted scores for some class, and showing only the best and worst matches [3, 26]. A more effective presentation technique evaluated by Amershi *et al.* consists of summarizing model quality while presenting low-certainty samples [5].

2.2 Text Applications

As the vast number of digital documents continues to increase, automated text categorization, information extraction, and summarization have witnessed particular interest in the context of ML.

Abstrackr is a standalone annotation tool independent of its ML components aiming to semi-automate the laborious task of citation screening for systematic reviews in clinical research settings [63]. The user screens documents arranged by an Active Learning ordering function, manually accepting or rejecting individual citations while entering additional relevant terms. Terms indicated as relevant or irrelevant by the user appear highlighted in differing colors within the text. Highlighting words or n-grams appears to maximize user perception of the features being exploited by the model and improve the understanding of the underlying function, including its deficiencies [22].

ML is especially useful when data is large and complex, and the visualizations and interactions provided in Interactive ML applications should account for volume and dimensionality. Visualizations like Word Tree and DocuBurst employ interactive layouts to reflect semantic content and enable rapid querying and exploration of bodies of text [18, 66]. The cognitive advantages of spatial representations of information are well documented and can effectively support Interactive ML applications, as seen in iVizTRANS and NEREx; two interactive visual analytics tools used to iteratively train ML classifiers on transportation data and conversation transcripts, respectively [7, 23, 24, 69].

A different approach is taken by Podium, a prototype system enabling non-expert users to rank multi-variate data points by dragging single rows in a table [62]. Similarly, the prototype BrainCel features a spreadsheet where the user can select which points to edit, add to the training set, or predict [54]. In a user study, the cycle of editing, learning and guessing within the table successfully encouraged participants to improve the model. Despite the lack of tables or spreadsheets as an interactive or visualization technique in text applications, the documented success of simple interfaces in enabling non-expert users to build ML models suggests a promising avenue [53].

Given the significant size of qualitative data sets and the time-consuming and laborious nature of coding, several attempts have been made to implement NLP techniques and ML models to support qualitative researchers [19, 20, 29, 39, 41, 46, 61]. Ranging from automatic content analysis to automatic coding, relevant work reveals low accuracy as the primary limitation of these systems. The tendency to advocate for a hybrid approach is commonly justified by the inadequacy of "one-size-fits-all" models to capture contextual nuance. An additional range of limitations discussed by Chen *et al.*, such as a lack of understanding between disciplines, points to Interactive ML techniques as possible solutions [16].

Recent work on AI-assisted data annotation presented and evaluated PaTAT, a human-AI collaborative tool that assists users with qualitative coding by implementing explainable interactive pattern synthesis to provide coding suggestions in the initial phase of the analysis [28]. The authors stress that, while in most domains, Interactive ML systems focus primarily on the optimization of the model, in domains such as QDA, scaffolding human learning is just as if not more important. After all, qualitative analysis is creative, reflexive and subjective [13], and entails the iterative exploration and review of new or existing patterns [12].

2.3 Research Gap and Contribution

Compared to previous work, our primary contribution is an improved understanding of how non-ML experts engage with and experience an Interactive ML system. The research gap we address is centered on the limited exploration of non-expert interactions with Interactive ML tools and the misconceptions and biases they may hold. Previous research has proposed and refined design principles, highlighted the technical benefits of Interactive ML, and evaluated prototypes, but the end-user experience remains under-explored. This gap is significant because it affects the design and deployment of user-friendly ML systems that can democratize access to advanced ML-driven tools by removing the need for experts to refine models and achieve greater performance.

3 QDA AS AN APPLICATION AREA FOR INTERACTIVE ML

The literature on Interactive ML identifies one of the greatest advantage in the ability for non-experts in ML to drive model refinement through low-cost trial and error or focused experimentation with inputs and outputs [3]. Applications generally assume a considerable degree of domain knowledge from the end-user, since overall familiarity with the data is required for accurate model inspection and feedback assignment. Therefore, we started identifying use cases for a system that would allow users to use their own data to achieve obtainable and personally useful goals.

Prior work highlighted the potential to apply ML to QDA [16, 28]. However, progress in applying ML to social science research has been relatively slow compared to domains like medicine, as low accuracy has been generally identified as the main limitation of systems automating QDA [38]. We approached this issue believing that Interactive ML techniques might at least mitigate the resulting loss in system accuracy. Numerous applications implementing the Interactive ML cycle have been evaluated in user studies involving non-expert participants, demonstrating that efficient feedback assignment and model inspection techniques are sufficient in building accurate models [23, 28, 54, 62, 63, 69].

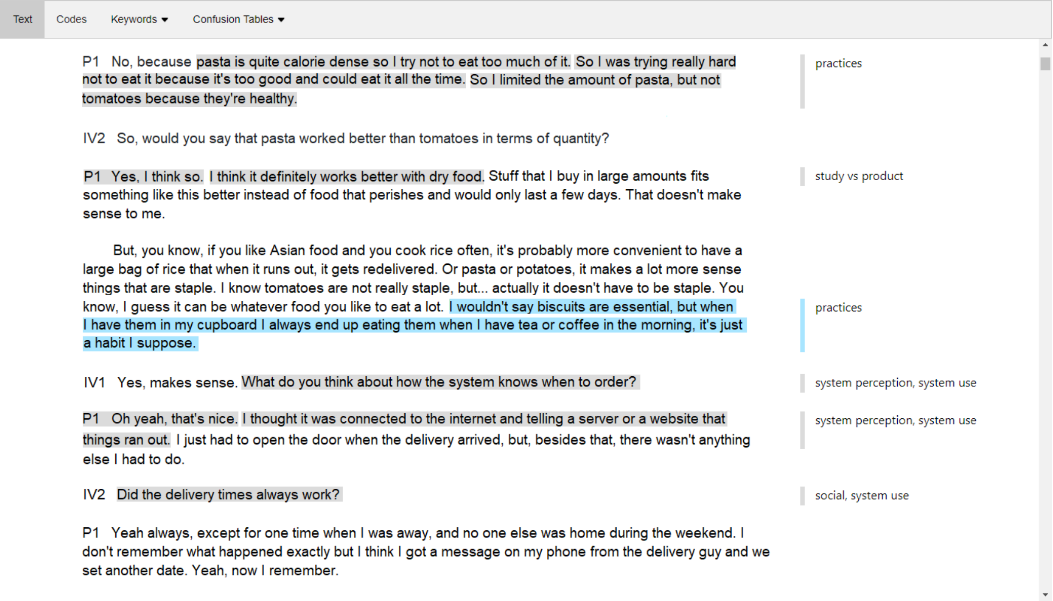
An additional issue in applying conventional ML to QDA is that building a learning model is not the primary goal of the social scientist. While ML models require a large quantity of classified data under predefined classes, new categories and concepts are likely to emerge during the coding phase, some of which might appear very infrequently. This, combined with calls from the literature to enhance, rather than supplant, the work of human coders [39], prompted us to consider a different approach to code automation. Instead of automating the coding process, we saw an opportunity to assist researchers in reflecting on their completed analysis by providing additional automated coding suggestions.

4 TACA: THEMATIC ANALYSIS CODING ASSISTANT

To enable a user study around the application of Interactive ML to qualitative thematic analysis, we developed TACA: an advanced, fully functioning GUI desktop application to assist the coding phase of the analysis. After users have performed at least an initial manual pass of the analysis, they can import the coded data set into TACA, which then trains an ML classifier to suggest how such initial analysis could be extended by assigning the user-defined themes to additional sentences that were not previously coded. Users can then inspect the output of the classifier (i.e., the coding suggestions), consequently modify the training data (i.e., re-labeling sentences from one theme to another), re-train the ML classifier to interactively refine it and, in so doing, customize it to produce more accurate coding suggestions.

Because qualitative data can often be confidential, and researchers might not have had permission to share it, it was critical to design and implement TACA as a stand-alone desktop application that could be used offline (i.e., without any data being transferred over the Internet, so no server support). Designed to support different software and strategies, TACA allows users to import the coded text and select either Microsoft Word or popular QDA software NVivo[1], MAXQDA[2] and Dedoose[3] as the original coding environment. After importing the data, users can define a list of terms to exclude from the analysis, such as transcript artifacts or additional stop words that might be specific to the data set that is being analyzed.

Following the setup, once the tool finishes extracting data, training the model, and classifying new sentences, the user is presented with the Text page, containing the entire transcript with the coded sentences. Highlighted in gray are the user-coded sentences, while those predicted by the model appear in blue, seen in Figure 1. Theme names appear in line with the respective sentences, in a similar fashion to comments in Microsoft Word and NVivo, and are also shown in a tooltip on mouseover.



P1 No, because pasta is quite calorie dense so I try not to eat too much of it. So I was trying really hard not to eat it because it's too good and could eat it all the time. So I limited the amount of pasta, but not tomatoes because they're healthy. — practices

IV2 So, would you say that pasta worked better than tomatoes in terms of quantity?

P1 Yes, I think so. I think it definitely works better with dry food. Stuff that I buy in large amounts fits something like this better instead of food that perishes and would only last a few days. That doesn't make sense to me. — study vs product

But, you know, if you like Asian food and you cook rice often, it's probably more convenient to have a large bag of rice that when it runs out, it gets redelivered. Or pasta or potatoes, it makes a lot more sense things that are staple. I know tomatoes are not really staple, but... actually it doesn't have to be staple. You know, I guess it can be whatever food you like to eat a lot. I wouldn't say biscuits are essential, but when I have them in my cupboard I always end up eating them when I have tea or coffee in the morning, it's just a habit I suppose. — practices

IV1 Yes, makes sense. What do you think about how the system knows when to order? — system perception, system use

P1 Oh yeah, that's nice. I thought it was connected to the internet and telling a server or a website that things ran out. I just had to open the door when the delivery arrived, but, besides that, there wasn't anything else I had to do. — system perception, system use

IV2 Did the delivery times always work? — social, system use

P1 Yeah always, except for one time when I was away, and no one else was home during the weekend. I don't remember what happened exactly but I think I got a message on my phone from the delivery guy and we set another date. Yeah, now I remember.

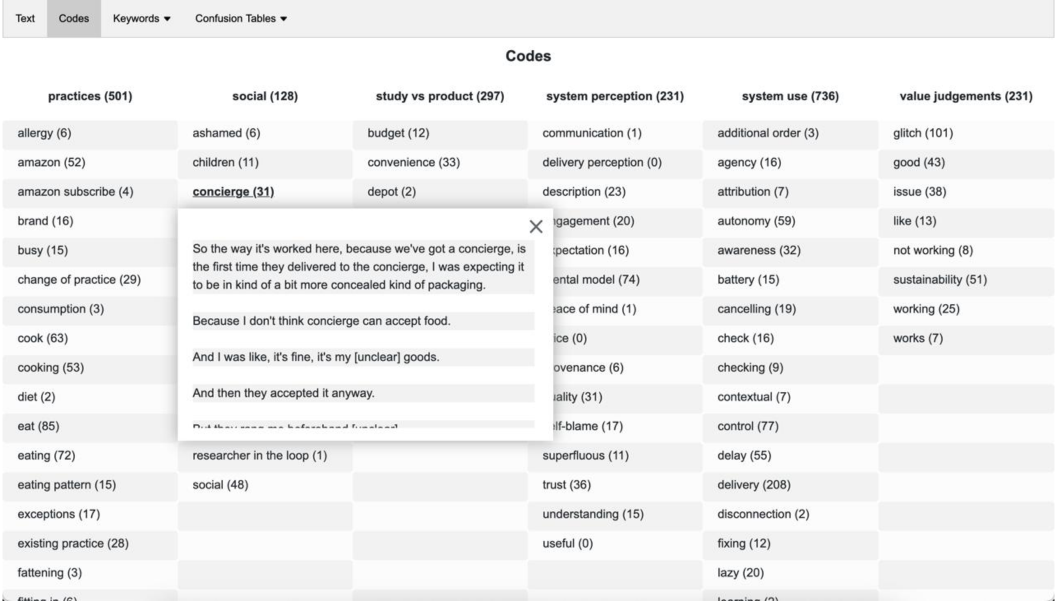
Fig. 1. Text page showing highlighted user-coded sentences in gray and classified sentences in blue.

A navigation bar at the top provides links to three other pages: Codes, Keywords, and Confusion Tables. The Codes page contains a basic lookup table for user-defined codes and respective themes. Seen in Figure 2, this page is included to allow the user to revisit their manual coding by clicking on any code to see all the associated sentences.

The Keywords tab includes a drop-down menu containing three pages: Train Keywords, Predict Keywords, and All Keywords. In line with prior research [22], we hypothesize that giving salience

[1] https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home
[2] https://www.maxqda.com/
[3] https://www.dedoose.com/

to indicative keywords accelerates the assessment process, and sort terms by frequency where each column is a theme, shown in Figure 3. In the Train Keywords table, the tool extracts all individual words that were manually coded under each theme by the user. In the Predict Keywords table, the words are only extracted from the classifications of the model. The All Keywords table is a combination of both Train and Predict. In all tables, a frequency counter is displayed next to each word, indicating the number of sentences that contain it. Each word can be clicked to reveal the list of sentences, individually highlighted in gray for training samples and blue for predicted. Unique to all Keywords Tables is the re-labeling interaction that allows users to drag and drop either a keyword or a single sentence from one column to the other, or to a bin. We exploit keywords as handles for groups of sentences to enable the user to re-label multiple data points, or sentences, at once.



Codes

practices (501)	social (128)	study vs product (297)	system perception (231)	system use (736)	value judgements (231)
allergy (6)	ashamed (6)	budget (12)	communication (1)	additional order (3)	glitch (101)
amazon (52)	children (11)	convenience (33)	delivery perception (0)	agency (16)	good (43)
amazon subscribe (4)	**concierge (31)**	depot (2)	description (23)	attribution (7)	issue (38)
brand (16)			gagement (20)	autonomy (59)	like (13)
busy (15)			pectation (16)	awareness (32)	not working (8)
change of practice (29)			ental model (74)	battery (15)	sustainability (51)
consumption (3)			ace of mind (1)	cancelling (19)	working (25)
cook (63)			ice (0)	check (16)	works (7)
cooking (53)			ovenance (6)	checking (9)	
diet (2)			ality (31)	contextual (7)	
eat (85)			lf-blame (17)	control (77)	
eating (72)	researcher in the loop (1)		superfluous (11)	delay (55)	
eating pattern (15)	social (48)		trust (36)	delivery (208)	
exceptions (17)			understanding (15)	disconnection (2)	
existing practice (28)			useful (0)	fixing (12)	
fattening (3)				lazy (20)	
fitting in (6)				learning (2)	

So the way it's worked here, because we've got a concierge, is the first time they delivered to the concierge, I was expecting it to be in kind of a bit more concealed kind of packaging.

Because I don't think concierge can accept food.

And I was like, it's fine, it's my [unclear] goods.

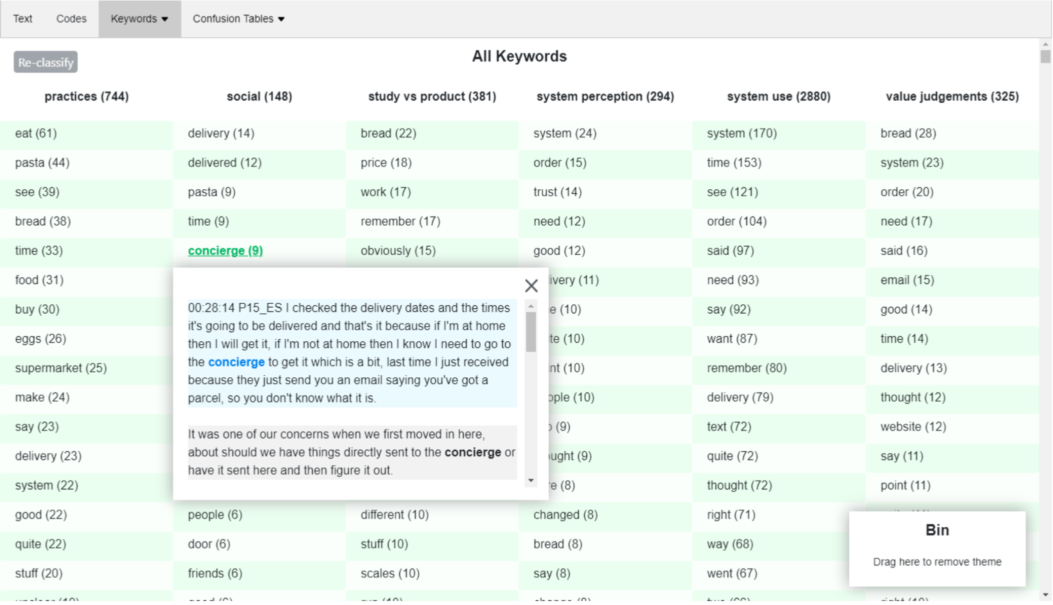
And then they accepted it anyway.

But they rang me beforehand [unclear]

Fig. 2. Codes page showing a lookup table for user-defined codes and respective themes.

After interacting with the table, the button Re-classify can be clicked to re-train the classifier. Once the re-classification ends (generally taking from 10-20 seconds to a few minutes, depending on the data and the computer speed) and the table is updated, individual cells where the frequency changed by more than half its original value are highlighted, following the design guidelines for dynamic visualizations in progressive analysis by Stolper et al. [59]. Additionally, because of the non-deterministic nature of the gradient boosting classifier, highlighting serves to suggest which changes are most likely due to re-training. Because each re-labeled sentence can propagate changes to other parts of the Keywords Table, the same technique is also employed after each drag-and-drop interaction.

In the final page, the Confusion Tables display confusion matrices for each theme in a table. Shown in Figure 4, each column contains true/false positive/negative samples, represented as keywords in the same way as in the Keywords Table. Aiming to facilitate the assessment of the current model state, keywords can be clicked to reveal the respective sentences.



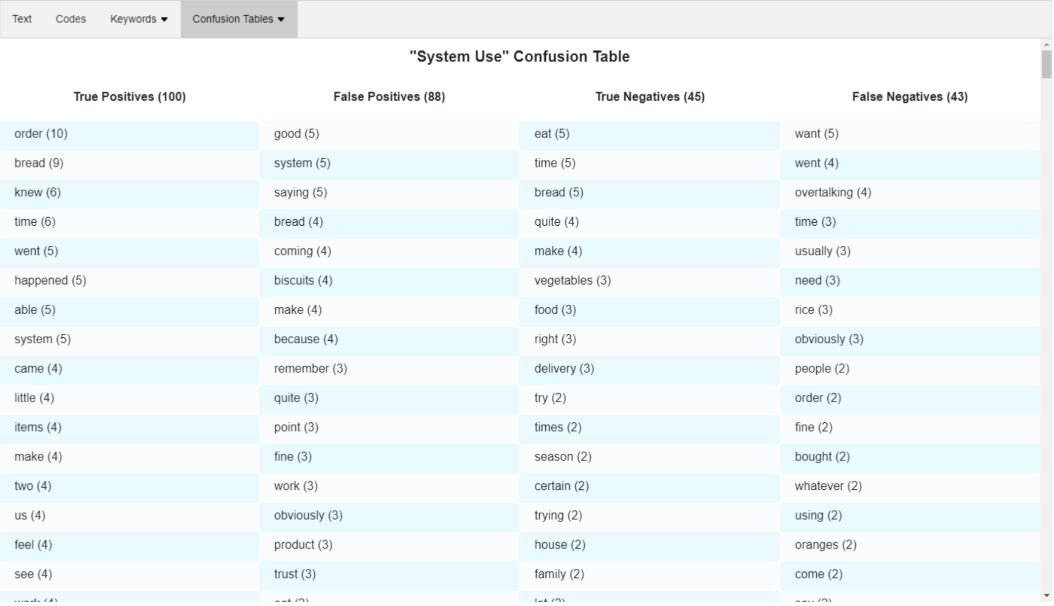
All Keywords

practices (744)	social (148)	study vs product (381)	system perception (294)	system use (2880)	value judgements (325)
eat (61)	delivery (14)	bread (22)	system (24)	system (170)	bread (28)
pasta (44)	delivered (12)	price (18)	order (15)	time (153)	system (23)
see (39)	pasta (9)	work (17)	trust (14)	see (121)	order (20)
bread (38)	time (9)	remember (17)	need (12)	order (104)	need (17)
time (33)	concierge (9)	obviously (15)	good (12)	said (97)	said (16)
food (31)			ivery (11)	need (93)	email (15)
buy (30)			e (10)	say (92)	good (14)
eggs (26)			te (10)	want (87)	time (14)
supermarket (25)			nt (10)	remember (80)	delivery (13)
make (24)			ople (10)	delivery (79)	thought (12)
say (23)			(9)	text (72)	website (12)
delivery (23)			ught (9)	quite (72)	say (11)
system (22)			e (8)	thought (72)	point (11)
good (22)	people (6)	different (10)	changed (8)	right (71)	
quite (22)	door (6)	stuff (10)	bread (8)	way (68)	
stuff (20)	friends (6)	scales (10)	say (8)	went (67)	

00:28:14 P15_ES I checked the delivery dates and the times it's going to be delivered and that's it because if I'm at home then I will get it, if I'm not at home then I know I need to go to the concierge to get it which is a bit, last time I just received because they just send you an email saying you've got a parcel, so you don't know what it is.

It was one of our concerns when we first moved in here, about should we have things directly sent to the **concierge** or have it sent here and then figure it out.

Fig. 3. All Keywords Table page showing the most frequently occurring terms for each theme.



"System Use" Confusion Table

True Positives (100)	False Positives (88)	True Negatives (45)	False Negatives (43)
order (10)	good (5)	eat (5)	want (5)
bread (9)	system (5)	time (5)	went (4)
knew (6)	saying (5)	bread (5)	overtalking (4)
time (6)	bread (4)	quite (4)	time (3)
went (5)	coming (4)	make (4)	usually (3)
happened (5)	biscuits (4)	vegetables (3)	need (3)
able (5)	make (4)	food (3)	rice (3)
system (5)	because (4)	right (3)	obviously (3)
came (4)	remember (3)	delivery (3)	people (2)
little (4)	quite (3)	try (2)	order (2)
items (4)	point (3)	times (2)	fine (2)
make (4)	fine (3)	season (2)	bought (2)
two (4)	work (3)	certain (2)	whatever (2)
us (4)	obviously (3)	trying (2)	using (2)
feel (4)	product (3)	house (2)	oranges (2)
see (4)	trust (3)	family (2)	come (2)

Fig. 4. Confusion Table page showing the most frequently occurring terms for each confusion matrix quadrant of the selected theme.

4.1 Implementation Details

TACA was implemented mostly in Python to leverage the availability of ML libraries. The PyQt[4] framework was used in conjunction with HTML and JavaScript for the UI. In terms of text processing, the transcript is segmented into sentences using the natural language processing library NLTK[5], and stop words defined in the same library are excluded. A vector is then generated for each sentence as the arithmetic mean of the embedding vectors representing each word in the sentence. The word embeddings are 50-dimensional and generated using the GloVe learning algorithm pre-trained on a generic Twitter data set[6]. Vectors corresponding to sentences that were coded by the user are associated to the corresponding codes and themes and used as training data. The vectors are then used to train a gradient boosting classifier XGBoost[7] to predict coding suggestions for uncoded sentences.

Due to the multi-label nature of the classification problem given that one sentence can belong to more than one theme, we use ClassifierChain from scikit-learn[8] and create a voting ensemble by arranging the binary XGBoost classifiers, one for each theme, into 10 chains in different, random orders. To address a possible imbalance in class distribution, we use MLSMOTE, a popular data augmentation algorithm for multi-label classification [15], before training the chains of classifiers on the labeled embeddings and predicting all the uncoded sentences in the text with a confidence threshold of 95%, taking the average of the binary predictions of the chains.

When the user imports the text containing coded sentences, the tool splits the data set into a training set and a test set using an 80:20 ratio to train the model on 80% of the coded sentences and generate the Confusion Tables on the remaining 20%. The process starts automatically after the end of the setup page. Due to computational constraints (the tool should run offline on as many personal computers as possible), the use of cross-validation was limited to an initial hyperparameter search for XGBoost using a collection of restaurant reviews coded by the researchers using the F1 score as the model validation metric. ML concepts such as the training/validation/test split, input features, algorithms, and hyperparameters are *not* presented to users as Confusion Tables were sufficiently advanced ML concepts for non-experts [57]. Additionally, TACA does not handle ambiguous data explicitly the way previous research in noisy data in ML proposed [31, 50, 55], because the automatic detection of incorrect samples in the training data set was infeasible due to the lack of ground truth in qualitative data. TACA was developed to process ambiguous data in terms of the subjectivity involved during the manual labeling process, as well as reviewing the coding suggestions generated by the classifier.

Multithreading enables TACA to load every page independently and simultaneously while prioritizing the currently selected page to reduce loading times. In all the Keywords pages, the button Re-classify creates a new training data set including the re-labeling changes from the user on training sentences or predicted sentences, or both. The new data set is used to train the same classifier again and generate new classifications, before updating every page in TACA. We release the code as open source[9].

[4] https://github.com/pyqt
[5] https://www.nltk.org/
[6] https://nlp.stanford.edu/projects/glove
[7] https://xgboost.readthedocs.io/
[8] https://scikit-learn.org/stable/modules/generated/sklearn.multioutput.ClassifierChain.html
[9] https://github.com/fmilana/taca

5 STUDY

The study was reviewed and approved by the University College London Interaction Centre (UCLIC) Ethics Committee (ID: UCLIC_2022_004_costanza). All participants were volunteers and provided informed consent before taking part in the study.

5.1 Participants

We recruited 20 participants from *Prolific*[10] (an online crowd-sourcing recruitment platform), a psychology and language science participant pool at our university, and among fellow researchers from different departments in other universities. The criteria set for recruitment were: minimum age of 18, fluency in English, at least 1 year of experience in QDA, and no experience in ML. Participant information is reported in Table 1. We acknowledge the gender imbalance among participants, but this reflects the demographic composition of the fields from which we recruited and where qualitative research is typically involved. Previous research in psychology and social sciences has found that gender can sometimes influence perspectives, cognitive styles, and analytical approaches [1, 58]. While the primary objective of our research was to expose a representative sample of QDA practitioners to Interactive ML, a promising future direction is to conduct more gender-balanced and gender-focused user studies in this area.

5.2 Procedure

We distributed TACA to our participants and asked them to install and run on their personal computers. A study information sheet provided instructions to import the transcript, and a description of all the pages, including the interaction with the Keywords Tables and the definitions of the terms used in the Confusion Tables (see Appendix B). Participants were instructed to use the tool until no more perceived value was gained, or after 20 minutes of use, whichever point was reached first.

5 of the 20 participants used their transcripts coded either in Microsoft Word, NVivo, MAXQDA or Dedoose. These transcripts were from studies participants conducted and were already analyzed for publication, ranging from a few months to a few years prior to the study. To facilitate recruitment, we also distributed a collection of 21 reviews of restaurants published in the newspaper *The Guardian*[11] between 2022 and 2023 to manually code by participants who did not have their own data sets available for the study. We chose restaurant reviews because the topic did not require specialized knowledge, and we expected reviews to be diverse yet having common themes. 21 reviews was the minimum length of the total text (25,000 words) at which TACA performed acceptably according to initial tests. Participants were instructed to analyze the reviews to identify 4 to 6 themes but were not provided a code book, so they were free to use either a deductive or inductive thematic analysis approach.

User interactions with the interface of TACA were timestamped and logged in a text file stored locally. The logged interactions included: launching and closing the tool, loading and switching pages, clicking on keywords to reveal the tooltip, closing the tooltip, dragging keywords or sentences noting their position in the table, and re-training the model. Participants were instructed to inspect the log text file, and, if satisfied that it did not contain any sensitive information, share it with the research team (all participants did).

Participants took part in a follow-up 20-minute, semi-structured interview focused on the experience of using TACA, including the general understanding of the tool and specific features (see Appendix A). Participants were asked to have the tool open on their machines during the interview, so that they could refer to the UI elements when answering questions, and so that video

[10] https://www.prolific.co/

[11] https://www.theguardian.com/food/restaurants+tone/reviews

Table 1. Participants information.

ID	Age	Sex	Occupation	Field of study/research	QDA experience (years)	Data used	Recruited from
P1	30-39	F	Postdoctoral researcher	HCI	3+	Own data	University
P2	30-39	F	PhD student	HCI	3+	Own data	University
P3	30-39	F	Postdoctoral researcher	Medicine	3+	Own data	University
P4	30-39	M	Postdoctoral researcher	HCI	3+	Own data	University
P5	27	F	Postdoctoral researcher	HCI	2	Own data	University
P6	20-29	F	Undergraduate student	Psychology	1	Restaurant reviews	Participant pool
P7	20-29	F	Undergraduate student	Psychology	1	Restaurant reviews	Participant pool
P8	20-29	F	Undergraduate student	Social sciences	1	Restaurant reviews	Participant pool
P9	20-29	M	Undergraduate student	Economics	1	Restaurant reviews	Participant pool
P10	24	F	Undisclosed	Psychology	3+	Restaurant reviews	Prolific
P11	20-29	F	Undergraduate student	Psychology	1	Restaurant reviews	Participant pool
P12	30-39	M	Postdoctoral researcher	HCI	3+	Restaurant reviews	University
P13	27	F	Undisclosed	Psychology	3+	Restaurant reviews	Prolific
P14	20-29	F	Unemployed	English literature	1	Restaurant reviews	Prolific
P15	20-29	F	Undergraduate student	Psychology	2	Restaurant reviews	Participant pool
P16	30	F	Undisclosed	Psychology	3+	Restaurant reviews	Prolific
P17	20-29	M	Undergraduate student	Psychology	1	Restaurant reviews	Participant pool
P18	20-29	F	Undergraduate student	Psychology	1	Restaurant reviews	Participant pool
P19	20-29	F	Postgraduate student	Social sciences	3+	Restaurant reviews	Participant pool
P20	20-29	F	Postgraduate student	Digital humanities	3+	Restaurant reviews	Participant pool

recordings of their screen could be revisited later to contextualize parts of the interviews (just for those who worked on the restaurant reviews).

Participants who used their own data set were financially compensated with £10 for a total of 1 hour spent on the study. Those who used the restaurant reviews received £45 to account for the additional time spent coding, which made the study duration about 4.5 hours in total. These participants were free to spread the study engagement over multiple days, and all of them did over a period of 5-10 days.

5.3 Analysis

Following an inductive orientation where coding and theme development was driven by the data, the analysis aimed to investigate participants' own perspective and understanding of Interactive ML and TACA. Audio recordings from the interviews were transcribed verbatim and then analyzed using reflexive thematic analysis [13, 14] by the authors. At an early phase, the authors focused on the explicit meaning of participants' accounts by familiarizing themselves with the interview data. Next, initial codes were drawn from the interviews using manual line-by-line coding and over-arching themes were developed. A second coding iteration followed a discussion among the authors, who revisited the initial themes and modified them based on the new codes. The coding process was repeated a third time to ensure that codes were relevant and consistent throughout the transcripts, resulting in a total of 106 codes grouped into 5 themes, discussed in the following section.

6 FINDINGS

We report findings from semi-structured interviews and present situated data on system usage based on automatic interaction logs. Participants reported the value of an Interactive ML assistant, critical reflections on their thematic analysis, positivist thematic analysis views, misunderstanding of ML concepts, and personal blame for poor ML model performance.

6.1 System Usage from Automatic Interaction Logs

Participants spent, on average, 5:53 minutes in the Text page (*SD* = 3:56), 1:05 minutes in the Codes page (*SD* = 1:03), 1:23 minutes in the Train Keywords page (*SD* = 2:10), 7:00 minutes in the Predict Keywords page (*SD* = 5:59), 4:12 minutes in the All Keywords page (*SD* = 5:11), and 7:28 minutes in the Confusion tables (*SD* = 5:14). After re-training the model, participants stayed on the Keyword Tables 68% of the time and switched to the Text page 32% of the time. Confusion tables were only accessed subsequently, 26% of the times the model was re-trained.

The initial average F1 score of the multi-label classifier across participants was 0.58 (*SD* = 0.21), with a minimum score of 0.25 and a maximum score of 0.85. Out of the 20 participants, 12 re-trained the model at least once, and 5 re-trained it twice or more. Of the remaining 8 participants, 5 re-labeled at least one data point but did not re-train the model (some reported forgetting to press the re-train button), and 3 participants did neither. Of the 12 participants who re-trained the model at least once, 7 participants re-trained once, 1 participant re-trained twice, and 4 participants re-trained three times or more. The 17 participants who engaged in re-labeling did so, for 63% of the time, by dragging keywords (i.e., groups of sentences) instead of individual sentences. On average, these participants moved 6.1 keywords (*SD* = 9.3) and 3.5 single sentences (*SD* = 8.3). Of the 6.1 keywords, 3.7 (*SD* = 6.2) were moved without opening the tooltip revealing the list of sentences containing the word.

Participants who re-labeled at least one data point re-labeled, on average, 0.7 keywords (*SD* = 1.7) in the Train Keywords Table (i.e., after seeing the ML output they modified their own classification

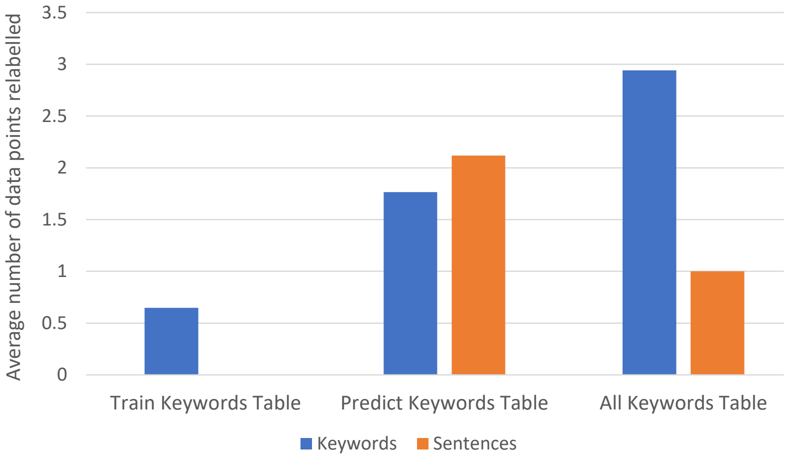


Fig. 5. Average number of data points re-labeled.

of sentences into themes), 2.0 (*SD* = 3.5) in the Predict Keywords Table (i.e., they corrected the ML model's classification of sentences into themes), and 3.3 keywords (*SD* = 9.7) in the All Keywords Table (i.e., they moved sentences across themes regardless of whether they were classified by themselves or by the model). Comparatively, no single sentences were re-labeled in Train, 2.4 (*SD* = 8.2) in Predict, and 1.1 (*SD* = 2.7) in All (See Figure 5) . The average row number keywords were re-labeled was 7.3 in the Train Keywords Table (*SD* = 8.0), 22.3 in Predict (*SD* = 28.9), and 23.7 (*SD* = 17.7) in All. Single sentences were dragged from row number 57.8 (*SD* = 106.8) in Predict and 35.8 (*SD* = 74.6) in All.

Participants who used their own data moved, on average, more keywords (8.2, *SD* = 8.7), compared to participants who were given restaurant reviews (6.2, *SD* = 10.1). The largest portion of keywords moved by participants using their own data was from the Predict Keywords Table (4.8, *SD* = 10.2), while participants using restaurant reviews moved keywords in the All Keywords Table the most (3.2, *SD* = 10.5). No relationship was found between the number of re-labeled data points and participant demographics, i.e. age, sex, occupation, field of study/research and QDA experience.

6.2 Evaluation Strategies for Model Inspection and Reflections on ML Output

Following the initial quantitative analysis of interaction data, we explore the evaluation strategies participants employed, reflecting on the model's output and their own coding practices. When presented with the results of the model aggregated as keywords in the Keywords Tables, participants spontaneously employed exploratory strategies to critically analyze and reflect on their own coding in a variety of ways. One strategy that frequently emerged was to identify connections between keywords and themes *"to see what relations they have, and if that relation is obvious"* (P1).

Keywords in the Keywords Tables were also considered effective in extracting information and summarizing results by *"synthesizing"* (P19) a large amount of text, *"giving you very comprehensive results"* (P8) to *"easily conclude something while reading the keywords included"* (P7). The sorting of keywords by frequency was reported to be *"useful"* (P5) in *"giving you an idea of what is most common"* (P17), to *"look at things that are more salient"* (P6) and to *"know what kind of work pops out"* (P7).

However, not all participants found aggregating data by frequency-sorted keywords effective. The limitation of keywords most commonly reported by participants, who failed to extract information to critically analyze their own data, concerned the lack of meaningful and unexpected terms that appeared at the top of the tables. *"The problem with this is that often the most meaningful nouns are actually never the ones that you're not expecting and are never the ones that have the most frequency, because the ones that are more frequent you already know them"* (P4).

The output of the model presented as coding suggestions also encouraged participants to reflect on their data and coding, and participants reported experiencing increased self-awareness of their own data analysis practice and perspective. Acknowledging a text excerpt as accurately coded, P12 described their own coding as *"selective"* when reflecting on a specific example: *"That's an example of language description that I haven't coded, which it's then accurately chosen. So I guess I've been quite selective as well with the things that I chose."*

Participants further demonstrated reflexivity when interpreting the differences between the model's predictions and their own coding. One participant, after noticing that *"some of the [keywords] actually fit really well into that particular theme"*, began to question themselves: *"and then I thought, why didn't I include that?"* to quickly follow up with an explanation: *"OK, I didn't include it because it was part of a particular phase that I wasn't focusing on with the study"* to then acknowledge their personal influence on data collection and alternative interpretations: *"because I chose to focus on this, that's what participants talked about. But actually, it's interesting that this theme touches on additional aspects"* (P5).

Participants believed that an advantage of adopting the tool in an iterative process is to address *"the main difficulty with analyzing qualitative data"*: *"rethinking whether what I coded is right or wrong or whether I need to change themes"* as suggestions would help either identify new themes (e.g., *"maybe there's some new theme coming up"* (P8)), or organize *"better themes and sub-themes"* (P10) overall.

6.3 Benefits and Challenges of Data Aggregation

While individual strategies for using keywords for model inspection and output interpretation varied, common themes emerged regarding their use for model inspection and batch re-labeling. In addition to facilitating data exploration and coding review, keywords enabled participants to assess the accuracy of the model and its suggestions strategically. Most participants followed a top-down, column-by-column approach, comparing the meaning of each keyword to the one of the theme in order to *"try to understand how the machine is doing, how it predicts"* (P19). Consequently, these participants were able to draw general conclusions such as: *"it's been doing very well, because most of the things are under the right categories"* (P14). Participants were almost always able to perceive at least some improvement in terms of accuracy. For example, P12 explained that *"the process of doing that is improving it"* , P8 noted that *"it got definitely more precise"*, and P7 reported that *"it made better predictions"*.

Participants generally approached the interactive aspect of the system with caution, to avoid *"getting rid of stuff that was maybe useful"* (P15). Much like the process of assessing the model's accuracy, participants compared the meaning of keywords and sentences to their respective themes to identify the *"obvious"* (P2, P3) ones that *"should belong somewhere else"* (P1, P2).

Some participants found dragging keywords instead of single sentences *"easier"* (P19), *"intuitive"* (P17), *"convenient"* and *"comprehensive"* (P8), as they would *"not need to check the text"* (P19), and could *"just put all the relevant keywords to their themes to organize them better"* (P8). Others still preferred the granularity of dragging individual sentences, as they *"felt like moving the keyword was too big of a move"* (P5), especially when keywords were ambiguous and represented sentences that naturally belonged to different themes.

The drag-and-drop interactions revealed a significant misconception around keywords from a group of participants who believed that they were re-labeling the word itself rather than the sentences that contained it. These participants were *"surprised that sometimes the words seem pretty random"* (P8), and that *"if you move one keyword into the bin, you get rid of that sentence and all the keywords attached to it"* (P11). The perception that TACA worked at the level of keywords rather than sentences confirmed a mental model mismatch: *"it's analyzing keywords, since that's a big part. It's got a whole section with keywords"* (P15).

6.4 Perception of the ML Model

Having identified the specific misconceptions around the use of keywords and sentences in the batch re-labeling process, we now shift to the broader perceptions of the ML model itself and how these shaped participants' experience using the system. Most participants clearly understood that the model was *"based on your previously trained data set"*, *"patterns"* (P20), and *"style of categorizing the codes"* (P9). Still, many participants viewed the model as an external source offering objective advice, a perspective reflected in numerous observations.. *"It's like an external source that's analyzing it in an objective manner in some way and telling you whether or not you got something right or wrong"* (P15). Partially, this was due to the perceived performance of the model, which was commonly overestimated (*"I don't see any inaccurate suggestions as far as I'm reading through. [...] I think it's brilliant!"* (P9)), but also to an underlying assumption that coding can be objectively correct or incorrect.

Confusion tables were introduced with the intention to enable model inspection, allowing non-expert users in ML to evaluate the performance of the model across each theme. The following exchange exemplifies the perception of Confusion Tables:

> Interviewer: *"These were ones that you did not code under 'privacy', but the model did."*
> P1: *"OK, so these could be the stuff that I might have missed then."*

The sentences shown under the false positives and false negatives columns were initially intended to allow participants to identify where the model failed. However, almost every participant seems to have considered them as indicating the accuracy of *their own* coding: *"OK, maybe I misread something or there is another interpretation. I think I looked at these more as suggestions"* (P2).

In some cases, the impression of the false positives and false negatives columns as suggestions developed only after comparing the outcome of the model to their own, differing classifications, and agreeing with the outcome of the model: *"I think I'm a bit conflicted because I came with the impression that it's a way for me to check if the model is performing well, but I misunderstood it, because now I'm the one that's left something out"* (P5). In other cases, this impression seemed to have arisen independently from exposure to situations where they concurred with the model's outcome. These participants were ready to question their own coding, but rarely the model's: *"The model must have some reasoning for categorizing these words into the false negatives."* (P9)

However, there was still value found in the Confusion Tables when evaluating the model. *"Forming an opinion depending on the quality of the false positives"* (P10) proved to be a popular approach: *"if you get a bunch of false positives, then that would mean that the things that were chosen from the program maybe shouldn't be as trusted and should be checked through"* (P15). Analyzing *"what [the model] is suggesting and maybe what it's also not suggesting"* was an effective strategy *"to see what the model thinks"* and determine: *"am I going to trust it? To what extent will I trust it?"* (P5).

6.5 Personal Blame for Poor Model Performance

The perception of an external source of objective advice naturally caused a second theme to emerge from the interviews: personal blame for poor model performance. Participants were able to use the

Text tab, Keywords Tables and Confusion Tables to detect instances where the model's performance was unexpected, evaluating the perceived accuracy of the suggestions by comparing them to their own coding. The consequence of the conflicting classifications was a widespread tendency to spontaneously attribute the cause of inaccuracy to a variety of factors that were exclusively traceable to the participants themselves, never to the quality of the model.

Participants (including those with greater experience in QDA) often mentioned their own lack of clarity in the themes and codes chosen: *"I might have included parts that aren't very useful to the specific theme that they fall under"* (P15), and *"I might have mixed some of the concepts"* (P16). *"It probably has to do with some error from my end"*, since *"the data set that I gave to the tool might have been a little bit at fault"* (P20): *"my themes weren't the clearest"* (P6) or *"not enough"* (P10).

From the understanding that the model was trained on their own data set, participants also inferred that *"the coding should be based on a large amount of data"* (P19). *"If I hadn't been coding much, then sometimes the results weren't what I expected because apparently the tool didn't have much to learn from"* (P7). Participants also frequently mentioned the ambiguous nature of qualitative data to justify the inaccurate suggestions given by the model: *"I feel like, if a word has different meanings, then that's where the confusion comes"* (P16).

The quality of the model was never questioned by any of the participants. Instead, the justifications to explain the inaccurate suggestions of the model were consistently unprompted, and given when participants were asked to identify situations where they believed the model performed inadequately.

6.6 Perceived and Anticipated Use of Interactive ML in QDA

Finally, we consider the wider context of integrating Interactive ML in existing data analysis workflows. Participants recognized that analyzing large quantities of text is time-consuming and welcomed the idea of implementing ML, acknowledging that TACA can *"take a lot of tedious work off your hands"* (P14) by accelerating the process of cross-checking for mistakes, identifying missed insights and nuances, reformulating codes and organizing ideas.

The desire to partially automate the coding phase was shared by many participants who envisioned an alternative use-case of the tool as one that could potentially save them even more time by *"not needing to code as many sentences, because it could predict my generating pattern and create codes based on my behavior"* (P9).

Nevertheless, there was a clearly perceived distinction between the researcher and the tool. *"I think the role of the tool is to organize or to scaffold the thinking of the researcher. It is a way for the researcher to test themselves and it could be quite helpful to mirror or reflect my processes as a researcher"* (P4). Participants recognized that, instead of replacing the researcher, TACA would complement them by cross-checking data, evaluating saturation, organizing existing ideas and identifying new insights.

Implicitly or explicitly, participants illustrated the potential influence of the tool on the manual coding phase of thematic analysis. Expecting the ML component of TACA to classify additional sentences for them, some participants who coded the restaurant reviews realized they *"could start becoming more lax with how thoroughly [they] coded everything"* towards the end of the text, since *"the AI has probably got enough information anyway"* (P17).

7 DISCUSSION

We have presented the design and evaluation of a thematic coding assistant. Through an analysis of interaction logs and semi-structured interviews, we have provided a situated account of how participants analyzed qualitative data using an Interactive ML system. Our findings demonstrate TACA as a functioning and usable tool to identify the benefits and challenges of enabling non-ML experts to engage with Interactive ML. We discuss how these have implications extending past the scope of our tool and can be applied to various domains outside QDA. The following three sections of the discussion focus on how Interactive ML supports reflexivity in data analysis, the tensions between the subjectivity of data and the expected objectivity of the ML model, and the general perception of ML driven by the experimental UI features we explored to facilitate the Interactive ML cycle.

7.1 Supporting Reflexivity with Interactive ML

Participants recognized and valued the process of reviewing their own analysis, identifying patterns, gaining deeper insights, and re-interpreting findings using TACA. The advantages of using Interactive ML in QDA reported by our participants confirm the results of Gebreegziabher et al. [28], which highlight the importance of the ability for researchers to refine and evolve their coding frameworks in collaboration with AI tools. Marathe et al. found that researchers desire automation only after having developed a codebook and coded a subset of data, particularly in extending their coding to unseen data [43], and most of the participants in our study confirmed this during the interviews. However, the benefits of Interactive ML extend beyond the automation and acceleration of data analysis.

More generally, participants also critically reflected on their own analysis after employing a variety of strategies to explore the ML output through the different parts of TACA. Previous research on Interactive ML states that result visualization techniques can enable users to assess the quality of the model and inform how to proceed in training [6]. Because our study involved subjective, ambiguous data with no objective ground truth, participants utilized result visualization to evaluate not only the performance of the model, but their own analysis too. These reflective practices were partly captured in the TACA interaction logs, which revealed that, in the Train Keywords Table, participants modified their own classification of sentences into themes.

In “Machine learners: Archaeology of a data practice”, Mackenzie argues that ML not only transforms the nature of knowledge but also impacts the practice of critical thoughts “as a mode of experimentation on one’s own conduct, thinking, and ways of being” [42]. During our interviews, many participants described the influence of their own presence and perspective as researchers on the findings when reflecting on and evaluating the differences between the model and their coding. This is crucial, since reflexivity is considered one of the pillars of critical research practices across various fields, including social sciences, humanities, and education [13, 27, 33, 36]. Reflexivity allows researchers to critically assess their own influence on the research process and outcomes, and in our study, was a reported benefit of evaluating the coding suggestions generated by the model.

It seems that reflexivity is driven by a tendency to justify the choices made during the manual coding phase of the analysis when faced with contrasting classifications from the model. Since most of the participants considered false positives and false negatives in the Confusion Table not as instances where the model failed, but sentences that they had possibly categorized under the wrong theme, recognizing their own perspective and possible bias towards the data was a direct consequence of questioning their own analysis.

Participants used TACA to reflect critically on their thematic analysis and often reassessed their own coding decisions when presented with the model’s suggestions. This behavior suggests that Interactive ML tools can foster a deeper engagement with data and encourage users to critically evaluate their own work. Reflexivity can be beneficial in various other fields where subjective interpretation is crucial. For example, in healthcare research, reflexivity can help medical professionals examine their diagnostic processes and treatment decisions, leading to more patient-centered care and improved health outcomes. Similarly, reflexivity can encourage researchers in cultural studies to examine their own biases and cultural assumptions, driving more nuanced and contextually rich analyses. Therefore, Interactive ML systems should be designed to promote critical engagement with data by providing clear and insightful feedback on both generated classifications and manually labeled data samples to allow for comparisons in a similar fashion to Confusion Tables in TACA.

7.2 Balancing Objectivity and Subjectivity in Interactive ML

Our study results emphasize reflexivity as a key benefit of Interactive ML, which is likely explained by the fact that most participants perceived the model as an *external, objective* source of advice, despite the subjective nature of the data involved. Rather than reviewing false positive and false negative samples as points where the model failed to classify their manually coded sentences, our participants often considered these belonging to an equally valid, if not better, interpretation of the data. This perception could also explain why participants re-labeled fewer data points than we expected and re-trained the model only a limited number of times.

A recent study by Yang et al. revealed that non-experts are generally more satisfied and trusting toward the outcome of ML compared to their professional counterparts [68], which can explain why participants almost always blamed themselves when recognizing that the model was performing poorly on their data set. In the specific context of QDA, a significant result of this perception is a shift towards a more positivist view. In the interviews, participants frequently mentioned the importance of subjectivity in their own analysis, but they just as often used terms like *"correct"*, *"incorrect"*, *"right"* or *"wrong"*, when evaluating their own coding or the output of the model. We suppose this could also have been influenced by the underlying goal of improving the accuracy of the model through the process of re-labeling and re-training, and the UI of TACA that displays terminology that are standard in ML, such as the Confusion Tables (with "true positives," "false negatives", etc.). We adopted such terminology because it is standard in ML, it is employed in various other fields, including inferential statistics and healthcare, and would be more accessible to non-experts compared to more complex measures, which can be overwhelming and misleading [10].

In fact, while recent work has found that non-experts often struggle with the standard terminologies and structural design of confusion matrices [57], most participants in our study clearly understood how to interpret the confusion tables and rarely required guidance during the interviews. Still, the terminology used might have inadvertently contributed to the perceived objectivity of the model, despite the fact that most participants recognized that the model was trained on their own, subjectively labeled data. In the Algorithmic Experience framework, algorithmic awareness refers to the users' understanding and knowledge of how algorithms function and impact their experience, and what influence the user can have on the results [2]. In our study, participants clearly recognized that re-labeling keywords was contributing to a closer alignment of the ML model to their interpretation of data, but they also demonstrated varying degrees of awareness regarding the ML processes embedded.

The quantification of data by qualitative researchers exposed to ML seems to be a pitfall that non-experts are commonly susceptible to, but the false perception of correctness seems prevalent in AI and extends beyond this group. The uncertainty, ambiguity and bias of ground truth data used to train ML models is rarely questioned, as highlighted by recent research observing how such data sets are constructed [45]. The reason might be that, in the ML academic communities, contributions are determined by the modeling work that takes place once the data is "cleaned". In reality, even within application domains where less subjectivity is at play, numerous external factors can significantly influence the process of data annotation [45]. Subjectivity in ML is also manifested in the processes of meaning-making, modeling choices, and data idiosyncrasies [35, 65], so, while our participants' perception that ML is intrinsically "objective" is not surprising, it should certainly be challenged.

Implementing explainability techniques in TACA was beyond the scope of this work, because we wanted to observe how participants would interact with the basic version of an Interactive ML tool without introducing additional complexity. However, our findings reveal a need for transparency in Interactive ML tools to help users understand the inherent limitations of ML models. Transparency has been found to encourage users to provide more labels [49] and with higher accuracy [52]. Explanations can mitigate the perception of an external, objective model and the consequent self-blame for errors by clearly communicating the probabilistic nature of ML classification. In fact, explanations have been found to increase user satisfaction with the output of the recommender [3] and, more notably, calibrate trust in the model, especially for non-experts [9, 22, 51]. Previous work has proposed guidelines and design implications for exposing Explainable AI to a general audience with the use of metaphors, visual aids, and interactive elements [56], and our results support the need of these recommendations for the use of explanations not just for ML practitioners in model diagnostics.

7.3 Understanding Perceptions of ML through UI Features

Building on the insights into the balance between objectivity and subjectivity in Interactive ML, we investigate in more depth how specific UI features influenced participants' perceptions and interactions with the model. Through the frequency-based keywords in the Confusion Tables and Keywords Tables we expected participants could get meaningful insights about the current state of the model, and also efficiently manipulate the large and high-dimensional data set for re-training the model, which is normally challenging. Our findings regarding keywords are specific to the UI of TACA and not universally applicable to all uses of Interactive ML. However, whether data aggregation techniques can facilitate model inspection and feedback assignment is an open question, and the design of similar techniques could benefit from the principles learned through our participants' experience with TACA.

We evaluated the presence of confusion matrices displayed as tables containing representative samples in the model inspection phase of Interactive ML. Participants spent, on average, around 6:30 minutes on these tables, switching to these pages after re-training the model most of the times. The results show that the identification of misclassifications in the inspection of the representative samples falling under "false positives" and "false negatives" can inform model evaluation by facilitating the semantic comparison between keyword and theme in a similar manner to the Keywords Tables. Our findings revealed that most participants naturally took different approaches to explore the output in the Keywords Tables. Most of our participants were confident in their assessment of the model after either comparing the meaning of each keyword to the theme or identifying semantic relations between the keywords in the same columns.

"Algorithmic control" in Algorithmic Experience refers to the ability of users to influence and and modify the behavior of algorithms to suit their needs and preferences [2]. Our study observed that participants actively engaged in activities that allowed them to re-classify data points and adjust model outputs based on their iterative feedback. A specific option for algorithmic user-control is to let users selectively turn off at least some data sources that are influencing the algorithm, and, in fact, a significant number of keywords were moved to the bin (to un-label groups of sentences).

The logged interactions revealed that participants preferred to re-label multiple samples simultaneously by dragging keywords rather than re-labeling individual sentences. Most of the keyword were re-labeled without revealing the list of associated sentences, since participants reported comparing the semantic meaning of each keyword directly to the allocated theme. These results suggest that interactions supporting simultaneously re-labeling multiple, semantically similar samples can be effective in the feedback assignment phase of the Interactive ML cycle, reducing the significant effort required to label data points in large data sets [30, 51, 67].

However, participants were generally hesitant to interact with the Keywords Tables and re-labeled fewer data points than we expected. Dragging and dropping whole keywords was considered by some too big of a move. In addition to enabling users to revert to a previous stage of the model, we hypothesize that anticipating the resulting changes before re-training could have increased the participants' confidence in re-labeling data points. Interfaces for feedback assignment in Interactive ML requires the most careful design in terms of both elements and interaction techniques [22], and visualizing anticipated changes can introduce transparency in the system, greatly affecting the quality of the response elicited from users [3].

The decision to use keywords as a data aggregation technique also introduced misconceptions around keywords and sentences. Some participants believed they were re-labeling individual words instead of the sentences that contained them, and that, consequently, TACA operated at the word level. We recognize this as a limitation of using keywords as handles, but also acknowledge an existing challenge of Interactive ML: users need to manipulate data, but data sets are challenging to represent and summarize due to their size and dimensionality. We speculate that the limitation we observed in our research is not exclusive to text. Different types of data, including images, audio, and other high-dimensional data forms are likely to present similar challenges when aggregation is used to facilitate model output inspection and batch re-labeling.

Other applications of ML on text, such as sentiment analysis and information retrieval, could benefit from aggregation to support the Interactive ML cycle. Our findings suggest that this approach can be effective, but it is essential to design around these features carefully to avoid misinterpretation. Interactive ML tools that aggregate data points should include complementary features that help users understand the relationship between grouped data and re-training the model. For example, visualizations that map aggregations to their corresponding set of individual data points could provide users with additional context. Also, systems could implement tooltips, detailed explanations, and interactive tutorials that guide users through the process of how data points are aggregated and the implications on the re-labeling process.

8 LIMITATIONS

Despite uncovering numerous insights into how users interact with Interactive ML systems, our study design introduced some limitations, which we discuss in this section.

One limitation relates to familiarity with the data. Interactive ML applications often assume that users possess domain knowledge, which is crucial for accurate model inspection and feedback [3]. Our intention was to recruit only participants who had their own coded data to analyze, but due to recruitment challenges, we eventually decided to provide newspaper restaurant reviews to those who did not have any data available. Eventually, only 5 out of 20 participants analyzed their own transcripts. Compared to this group, the other participants likely had a more limited understanding of the data, which may have affected their ability to provide effective model feedback, and thus may not have engaged as critically with the output of the model.

Another limitation is that our participants naturally focused much of their attention on understanding how the tool works, how to use its features, and how to interpret the results, which likely constrained their ability to critically evaluate the analysis itself. The tool instructions provided before the study (see Appendix B) were a product of many iterations during the pilot studies to ensure clarity and ease of use, while also gently introducing non-expert users to basic ML concepts. Still, the UI of the tool is novel and quite different from any existing QDA software that participants might have been familiar with, requiring additional cognitive effort and inevitably diverting attention away from the analysis.

Additional limitations include important aspects of ML that were not explored in this study, such as the use of explainability techniques, like word importance heat maps. We did not evaluate these techniques as incorporating them would have introduced too much complexity into the study, potentially overwhelming participants and detracting from the focus on Interactive ML. However, these could have helped participants better understand the decision-making process of the model and reduced misconceptions.

The study also did not involve the use of the transformer architecture for text classification as we would have had to artificially limit the performance of the model to obtain a sufficient number of false positives and false negatives in the Confusion Tables to create the same opportunities for critical feedback and interaction. Transformers are considered state-of-the-art models for text classification and would have provided a setting that more accurately reflects real-world scenarios, with superior performance and more precise results. However, limiting the performance of the model for the sake of critical feedback would have contradicted the very nature of transformer models in real-world effectiveness. This issue might be less problematic in scenarios where participants work with much larger data sets, as the ability of the model to generate false positives and false negatives naturally increases with more data.

9 CONCLUSION

This paper reported on a user study where 20 participants without prior ML experience used TACA, a novel Interactive ML application designed and developed to enable the study. We focused on thematic analysis as a practical application of Interactive ML: thematic analysis involves individual interpretation of ambiguous data and hence it is suited for and can benefit from the iterative customization of models supported by Interactive ML. Our participants had at least one year of experience in thematic analysis, and used TACA to refine the analysis of a data set from their own qualitative research or one we provided to them (newspaper restaurant reviews), if they did not have data available.

TACA was effective in exposing our participants to Interactive ML and apply it on their data. Participants recognized the value of incorporating Interactive ML in the thematic analysis workflow as the presence of coding suggestions encouraged a more critical analysis of data. Keywords and Confusion Tables, which were presented within the TACA UI, also supported the model inspection and feedback assignment phases of the Interactive ML cycle but introduced misconceptions around the mental model of the tool. Finally, our findings suggest that users with no experience in ML tend to perceive the model as an external, objective entity in the absence of ground truth, and consequently blame themselves when the model performs poorly.

We believe that Interactive ML has significant advantages over conventional ML, but that the success of this alternative approach is strongly dependent on our understanding of user perception and interaction with ML models. We hope that our work can serve as a practical example of a contribution facing this direction and stimulate further interest in this particular intersection between HCI and AI.

ACKNOWLEDGMENTS

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC, ref: EP/N509577/1). The study was approved by the UCLIC Ethics Committee (ID: UCLIC_2022 _004_costanza). We would like to thank all the participants of the study as well as the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper. Data URI: https://doi.org/10.5522/04/28182962.

REFERENCES

[1] Chaham Alalouch. 2021. Cognitive Styles, Gender, and Student Academic Performance in Engineering Education. *Education Sciences* 11, 9 (Sept. 2021), 502. https://doi.org/10.3390/educsci11090502 Number: 9 Publisher: Multidisciplinary Digital Publishing Institute.

[2] Oscar Alvarado and Annika Waern. 2018. Towards Algorithmic Experience: Initial Efforts for Social Media Contexts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173860

[3] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine* 35, 4 (Dec. 2014), 105–120. https://doi.org/10.1609/aimag.v35i4.2513

[4] Saleema Amershi, Max Chickering, Steven M. Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. ModelTracker: Redesigning Performance Analysis Tools for Machine Learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 337–346. https://doi.org/10.1145/2702123.2702509

[5] Saleema Amershi, James Fogarty, Ashish Kapoor, and Desney Tan. 2009. Overview based example selection in end user interactive concept learning. In *Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology (UIST '09)*. Association for Computing Machinery, New York, NY, USA, 247–256. https://doi.org/10.1145/1622176.1622222

[6] Saleema Amershi, James Fogarty, Ashish Kapoor, and Desney Tan. 2011. Effective End-User Interaction with Machine Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 25, 1 (Aug. 2011), 1529–1532. https://ojs.aaai.org/index.php/AAAI/article/view/7964

[7] Christopher Andrews, Alex Endert, and Chris North. 2010. Space to think: large high-resolution displays for sensemaking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. Association for Computing Machinery, New York, NY, USA, 55–64. https://doi.org/10.1145/1753326.1753336

[8] Dustin Arendt, Emily Saldanha, Ryan Wesslen, Svitlana Volkova, and Wenwen Dou. 2019. Towards rapid interactive machine learning: evaluating tradeoffs of classification without representation. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 591–602. https://doi.org/10.1145/3301275.3302280

[9] Amid Ayobi, Katarzyna Stawarz, Dmitri Katz, Paul Marshall, Taku Yamagata, Raul Santos-Rodríguez, Peter Flach, and Aisling Ann O'Kane. 2021. Machine Learning Explanations as Boundary Objects: How AI Researchers Explain and Non-Experts Perceive Machine Learning. In *2021 Joint ACM Conference on Intelligent User Interfaces Workshops, ACMIUI-WS 2021*. CEUR Workshop Proceedings.

[10] Emmanuelle Beauxis-Aussalet and Lynda Hardman. 2014. Visualization of Confusion Matrix for Non-Expert Users (Poster). (Oct. 2014). https://ir.cwi.nl/pub/22775

[11] Stuart Berg, Dominik Kutra, Thorben Kroeger, Christoph N. Straehle, Bernhard X. Kausler, Carsten Haubold, Martin Schiegg, Janez Ales, Thorsten Beier, Markus Rudy, Kemal Eren, Jaime I. Cervantes, Buote Xu, Fynn Beuttenmueller, Adrian Wolny, Chong Zhang, Ullrich Koethe, Fred A. Hamprecht, and Anna Kreshuk. 2019. ilastik: interactive machine learning for (bio)image analysis. *Nature Methods* 16, 12 (Dec. 2019), 1226–1232. https://doi.org/10.1038/s41592-019-0582-9

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

[13] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (Aug. 2019), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

[14] David Byrne. 2022. A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & Quantity* 56, 3 (June 2022), 1391–1412. https://doi.org/10.1007/s11135-021-01182-y

[15] Francisco Charte, Antonio J. Rivera, María J. del Jesus, and Francisco Herrera. 2015. MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems* 89 (2015), 385–397. https://doi.org/10.1016/j.knosys.2015.07.019

[16] 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 Transactions on Interactive Intelligent Systems* 8, 2 (June 2018), 9:1–9:20. https://doi.org/10.1145/3185515

[17] May Y. Choi and Christopher Ma. 2020. Making a big impact with small datasets using machine-learning approaches. *The Lancet Rheumatology* 2, 8 (Aug. 2020), e451–e452. https://doi.org/10.1016/S2665-9913(20)30217-4

[18] Christopher Collins, Sheelagh Carpendale, and Gerald Penn. 2009. DocuBurst: Visualizing Document Content using Language Structure. *Computer Graphics Forum* 28, 3 (2009), 1039–1046. https://doi.org/10.1111/j.1467-8659.2009.01439.x

[19] Kevin Crowston, Eileen E. Allen, and Robert Heckman. 2012. Using natural language processing technology for qualitative data analysis. *International Journal of Social Research Methodology* 15, 6 (Nov. 2012), 523–543. https:

//doi.org/10.1080/13645579.2011.625764

[20] Kevin Crowston, Xiaozhong Liu, and Eileen Allen. 2010. Machine Learning and Rule-Based Automated Coding of Qualitative Data. *Proceedings of the American Society for Information Science and Technology* 47 (Nov. 2010), 1–2. https://doi.org/10.1002/meet.14504701328

[21] André Dantas de Medeiros, Nayara Pereira Capobiango, José Maria da Silva, Laércio Junio da Silva, Clíssia Barboza da Silva, and Denise Cunha Fernandes dos Santos Dias. 2020. Interactive machine learning for soybean seed and seedling quality classification. *Scientific Reports* 10, 1 (July 2020), 11267. https://doi.org/10.1038/s41598-020-68273-y

[22] John J. Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning. *ACM Transactions on Interactive Intelligent Systems* 8, 2 (June 2018), 8:1–8:37. https://doi.org/10.1145/3185517

[23] Mennatallah El-Assady, Rita Sevastjanova, Bela Gipp, Daniel Keim, and Christopher Collins. 2017. NEREx: Named-Entity Relationship Exploration in Multi-Party Conversations. *Computer Graphics Forum* 36, 3 (2017), 213–225. https://doi.org/10.1111/cgf.13181

[24] Alex Endert, Patrick Fiaux, and Chris North. 2012. Semantic interaction for visual text analytics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. Association for Computing Machinery, New York, NY, USA, 473–482. https://doi.org/10.1145/2207676.2207741

[25] Jerry Alan Fails and Dan R. Olsen. 2003. Interactive machine learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces (IUI '03)*. Association for Computing Machinery, New York, NY, USA, 39–45. https://doi.org/10.1145/604045.604056

[26] James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. 2008. CueFlik: interactive concept learning in image search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. Association for Computing Machinery, New York, NY, USA, 29–38. https://doi.org/10.1145/1357054.1357061

[27] Joyce S. Fontana. 2004. A Methodology for Critical Science in Nursing. *Advances in Nursing Science* 27, 2 (June 2004), 93.

[28] Simret Araya Gebreegziabher, Zheng Zhang, Xiaohang Tang, Yihao Meng, Elena L. Glassman, and Toby Jia-Jun Li. 2023. PaTAT: Human-AI Collaborative Qualitative Coding with Explainable Interactive Rule Synthesis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3544548.3581352

[29] Justin Grimmer and Brandon M. Stewart. 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis* 21, 3 (2013), 267–297. https://doi.org/10.1093/pan/mps028

[30] Alex Groce, Todd Kulesza, Chaoqiang Zhang, Shalini Shamasunder, Margaret Burnett, Weng-Keen Wong, Simone Stumpf, Shubhomoy Das, Amber Shinsel, Forrest Bice, and Kevin McIntosh. 2014. You Are the Only Possible Oracle: Effective Test Selection for End Users of Interactive Machine Learning Systems. *IEEE Transactions on Software Engineering* 40, 3 (March 2014), 307–323. https://doi.org/10.1109/TSE.2013.59

[31] Shivani Gupta and Atul Gupta. 2019. Dealing with Noise Problem in Machine Learning Data-sets: A Systematic Review. *Procedia Computer Science* 161 (Jan. 2019), 466–474. https://doi.org/10.1016/j.procs.2019.11.146

[32] Björn Hartmann, Leith Abdulla, Manas Mittal, and Scott R. Klemmer. 2007. Authoring sensor-based interactions by demonstration with direct manipulation and pattern recognition. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. Association for Computing Machinery, New York, NY, USA, 145–154. https://doi.org/10.1145/1240624.1240646

[33] Andrew Holmes. 2020. Researcher Positionality - A Consideration of Its Influence and Place in Qualitative Research - A New Researcher Guide. *Shanlax International Journal of Education* 8 (Sept. 2020), 1–10. https://doi.org/10.34293/education.v8i4.3232

[34] Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics* 3, 2 (June 2016), 119–131. https://doi.org/10.1007/s40708-016-0042-6

[35] Saleha Javed, Tosin P. Adewumi, Foteini Simistira Liwicki, and Marcus Liwicki. 2021. Understanding the Role of Objectivity in Machine Learning and Research Evaluation. *Philosophies* 6, 1 (March 2021), 22. https://doi.org/10.3390/philosophies6010022

[36] Dev Jootun, Gerry McGhee, and Glenn R. Marland. 2009. Reflexivity: promoting rigour in qualitative research. *Nursing Standard* 23, 23 (Feb. 2009), 42–47.

[37] Jacob Kittley-Davies, Ahmed Alqaraawi, Rayoung Yang, Enrico Costanza, Alex Rogers, and Sebastian Stein. 2019. Evaluating the Effect of Feedback from Different Computer Vision Processing Stages: A Comparative Lab Study. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300273

[38] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. Computational Social Science. *Science* 323, 5915 (Feb. 2009), 721–723. https://doi.org/10.1126/science.1167742

[39] Seth C. Lewis, Rodrigo Zamith, and Alfred Hermida. 2013. Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods. *Journal of Broadcasting & Electronic Media* 57, 1 (Jan. 2013), 34–52. https://doi.org/10.1080/08838151.2012.761702

[40] Ching Ya Liao, Pangfeng Liu, and Jan-Jan Wu. 2020. Convolution Filter Pruning for Transfer Learning on Small Dataset. In *2020 International Computer Symposium (ICS)*. 79–84. https://doi.org/10.1109/ICS51289.2020.00025

[41] Jasy Liew, Nancy McCracken, Shichun Zhou, and Kevin Crowston. 2014. Optimizing Features in Active Machine Learning for Complex Qualitative Content Analysis. 44–48. https://doi.org/10.3115/v1/W14-2513

[42] Adrian Mackenzie. 2017. *Machine Learners: Archaeology of a Data Practice.* MIT Press.

[43] Megh Marathe and Kentaro Toyama. 2018. Semi-Automated Coding for Qualitative Research: A User-Centered Inquiry and Initial Prototypes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173922

[44] Georgios Mastorakis. 2018. Human-like machine learning: limitations and suggestions. *arXiv:1811.06052 [cs]* (Nov. 2018). http://arxiv.org/abs/1811.06052

[45] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 115:1–115:25. https://doi.org/10.1145/3415186

[46] 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)*. Association for Computing Machinery, New York, NY, USA, 3–8. https://doi.org/10.1145/2957276.2957280

[47] Van Hiep Phung and Eun Joo Rhee. 2019. A High-Accuracy Model Average Ensemble of Convolutional Neural Networks for Classification of Cloud Image Patches on Small Datasets. *Applied Sciences* 9, 21 (Jan. 2019), 4500. https://doi.org/10.3390/app9214500

[48] Reid Porter, James Theiler, and Don Hush. 2013. Interactive Machine Learning in Data Exploitation. *Computing in Science Engineering* 15, 5 (Sept. 2013), 12–20. https://doi.org/10.1109/MCSE.2013.74

[49] Al M. Rashid, Kimberly Ling, Regina D. Tassone, Paul Resnick, Robert Kraut, and John Riedl. 2006. Motivating participation by displaying the value of contribution. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. Association for Computing Machinery, New York, NY, USA, 955–958. https://doi.org/10.1145/1124772.1124915

[50] Veselin Raychev, Pavol Bielik, Martin Vechev, and Andreas Krause. 2016. Learning programs from noisy data. In *Proceedings of the 43rd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL '16)*. Association for Computing Machinery, New York, NY, USA, 761–774. https://doi.org/10.1145/2837614.2837671 event-place: St. Petersburg, FL, USA.

[51] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[52] Stephanie L. Rosenthal and Anind K. Dey. 2010. Towards maximizing the accuracy of human-labeled sensor data. In *Proceedings of the 15th International Conference on Intelligent User Interfaces (IUI '10)*. Association for Computing Machinery, New York, NY, USA, 259–268. https://doi.org/10.1145/1719970.1720006

[53] Advait Sarkar, Alan F Blackwell, Mateia Jamnik, and Martin Spott. 2014. Teach and try: A simple interaction technique for exploratory data modelling by end users. In *2014 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC'14)*. 53–56. https://doi.org/10.1109/VLHCC.2014.6883022

[54] Advait Sarkar, Mateja Jamnik, Alan F. Blackwell, and Martin Spott. 2015. Interactive visual machine learning in spreadsheets. In *2015 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC'15)*. 159–163. https://doi.org/10.1109/VLHCC.2015.7357211

[55] Jeffrey C. Schlimmer and Richard H. Granger. 1986. Incremental learning from noisy data. *Machine Learning* 1, 3 (Sept. 1986), 317–354. https://doi.org/10.1007/BF00116895

[56] Beatriz Severes, Carolina Carreira, Ana Beatriz Vieira, Eduardo Gomes, João Tiago Aparício, and Inês Pereira. 2023. The Human Side of XAI: Bridging the Gap between AI and Non-expert Audiences. In *Proceedings of the 41st ACM International Conference on Design of Communication (SIGDOC '23)*. Association for Computing Machinery, New York, NY, USA, 126–132. https://doi.org/10.1145/3615335.3623062 event-place: <conf-loc>, <city>Orlando</city>, <state>FL</state>, <country>USA</country>, </conf-loc>.

[57] Hong Shen, Haojian Jin, Ángel Alexander Cabrera, Adam Perer, Haiyi Zhu, and Jason I. Hong. 2020. Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 153:1–153:22. https://doi.org/10.1145/3415224

[58] Robert J. Sternberg and Li-fang Zhang. 2014. *Perspectives on Thinking, Learning, and Cognitive Styles.* Routledge. Google-Books-ID: YMeQAgAAQBAJ.

[59] Charles D. Stolper, Adam Perer, and David Gotz. 2014. Progressive Visual Analytics: User-Driven Visual Exploration of In-Progress Analytics. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (Dec. 2014), 1653–1662. https://doi.org/10.1109/TVCG.2014.2346574

[60] Simone Stumpf, Vidya Rajaram, Lida Li, Margaret Burnett, Thomas Dietterich, Erin Sullivan, Russell Drummond, and Jonathan Herlocker. 2007. Toward harnessing user feedback for machine learning. In *Proceedings of the 12th international conference on Intelligent user interfaces (IUI '07)*. Association for Computing Machinery, New York, NY, USA, 82–91. https://doi.org/10.1145/1216295.1216316

[61] P.J. Tierney. 2012. A qualitative analysis framework using natural language processing and graph theory. *The International Review of Research in Open and Distributed Learning* 13 (Nov. 2012), 173–189. https://doi.org/10.19173/irrodl.v13i5.1240

[62] Emily Wall, Subhajit Das, Ravish Chawla, Bharath Kalidindi, Eli T. Brown, and Alex Endert. 2018. Podium: Ranking Data Using Mixed-Initiative Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (Jan. 2018), 288–297. https://doi.org/10.1109/TVCG.2017.2745078

[63] Byron C. Wallace, Kevin Small, Carla E. Brodley, Joseph Lau, and Thomas A. Trikalinos. 2012. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium (IHI '12)*. Association for Computing Machinery, New York, NY, USA, 819–824. https://doi.org/10.1145/2110363.2110464

[64] Malcolm Ware, Eibe Frank, Geoffrey Holmes, Mark Hall, and Ian H Witten. 2001. Interactive machine learning: letting users build classifiers. *International Journal of Human-Computer Studies* 55, 3 (Sept. 2001), 281–292. https://doi.org/10.1006/ijhc.2001.0499

[65] Zeerak Waseem, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. 2021. Disembodied Machine Learning: On the Illusion of Objectivity in NLP. https://doi.org/10.48550/arXiv.2101.11974

[66] Martin Wattenberg and Fernanda B. Viégas. 2008. The Word Tree, an Interactive Visual Concordance. *IEEE Transactions on Visualization and Computer Graphics* 14, 6 (Nov. 2008), 1221–1228. https://doi.org/10.1109/TVCG.2008.172

[67] Weng-Keen Wong, Ian Oberst, Shubhomoy Das, Travis Moore, Simone Stumpf, Kevin McIntosh, and Margaret Burnett. 2011. End-user feature labeling: a locally-weighted regression approach. In *Proceedings of the 16th International Conference on Intelligent User Interfaces (IUI '11)*. Association for Computing Machinery, New York, NY, USA, 115–124. https://doi.org/10.1145/1943403.1943423

[68] Qian Yang, Jina Suh, Nan-Chen Chen, and Gonzalo Ramos. 2018. Grounding Interactive Machine Learning Tool Design in How Non-Experts Actually Build Models. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. ACM, Hong Kong China, 573–584. https://doi.org/10.1145/3196709.3196729

[69] Liang Yu, Wei Wu, Xiaohui Li, Guangxia Li, Wee Siong Ng, See-Kiong Ng, Zhongwen Huang, Anushiya Arunan, and Hui Min Watt. 2015. iVizTRANS: Interactive visual learning for home and work place detection from massive public transportation data. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST '15)*. 49–56. https://doi.org/10.1109/VAST.2015.7347630

[70] Ying Zhang and Chen Ling. 2018. A strategy to apply machine learning to small datasets in materials science. *npj Computational Materials* 4, 1 (May 2018), 1–8. https://doi.org/10.1038/s41524-018-0081-z

A SEMI-STRUCTURED INTERVIEW SCRIPT

Background and Experience

- Could you please describe your academic background and your experience with qualitative data analysis?
- How many years of experience do you have with qualitative analysis?

Coding Process

- How did the coding of the restaurant reviews go?
- Can you describe the transcript/project you used for this study?
- How long is the text?
- How long ago did you code your transcript?
- What software did you use to code your transcript? / the reviews?
- How long did it take in total?

- How many themes did you have?

Overall Experience

- How would you describe your overall experience using the tool?

Tool Functionality

- Describe what the tool does and how it works, like you would to a friend who has never seen it before?
- How would you explain how the tool does this?
- How does the tool make coding easier/harder/no difference?
- What do you think is the role of the researcher compared to the role of the tool?
- Clear separation QDA/assistant?

Accuracy of Suggestions

- From this tab, how accurate do you think the suggestions are compared to your own coding?
- Example of a suggestion that makes sense?
- Why do you think the sentence was suggested?/model was right?
- Example of a suggestion that does not make sense?
- Why do you think the sentence was suggested?/model was wrong?

Text Tab

- What did you think of the way the keywords are visualized in a table format and how they are sorted by frequency?
- Would you have sorted them differently?
- Have you ever seen your data like this?
- Have you opened the tooltip, and when do you think this would be most useful?
- Compared predict with train/all?
- What is the value (if any) of these tables in informing qualitative data analysis?

Reclassification Process

- Have you reclassified any keywords or sentences?
- Did dragging keywords make reclassifying easier or harder?
- Compared to single sentences?
- Give an example of when you used this feature (one or more examples)
- How did you decide which keywords or sentences to move before retraining the model?
- Average position of retrained keywords?
- Did you also reclassify trained samples, and if so, why?
- Example of when retraining the model gave results you expected/did not expect

Model Inspection

- How did you evaluate the current state of the model at each reclassification step?
- Where/how were you able to see which theme the model performed the best in?
- Which tabs did you switch to after reclassifying, and why?
- Do you feel the model has improved over the reclassifications, and how much at each step?
- What do you think the model/tool has learned based on the data and your interactions with it?
- What concepts emerged from the tables after each reclassification (in terms of what the model learned)?

Confusion Tables

- Do you remember the description of this tab from the instructions document?
- Have you ever heard of these terms before?
- Were the terms clear?
- What did you think about this table?
- Particular strategy around confusion tables (counters, false positive/negative columns, etc.)?
- How do you feel about situations where the model disagreed with your coding (false columns)?
- Tooltip?
- Which columns do you think are most useful for reflecting on your coding and why?
- Which columns do you think are most useful for evaluating the performance of the model and why?
- How would you use this table? (evaluate the model or for coding review?)

Features

- Most interesting or useful feature, (tool as part of research, writing a paper, etc)
- Least interesting or useful feature, (tool as part of research, writing a paper, etc)
- If you had to evaluate how well the tool is performing, which tab would you use and why?
- Do you think this tab is effective in the evaluation?
- Insights learned after using the tool besides sentences that you should have or should not have coded? Is there anything more general that emerged? Maybe a theme that you thought you could have added, changed, or removed?

Use Case

- Which data do you think is particularly suitable for TACA?
- Thoughts on using TACA different stages/projects (top down approach vs bottom up approach (no-go))?
- Would you see using the tool iteratively or just once?

Challenges and Benefits

- What are the challenges/limitations of coding data with existing tools?
- Are there any features that [chosen QDA] could benefit from any features from TACA?
- Which ones and why?
- Would TACA benefit from any features from [chosen QDA]?
- Which ones and why?

General Comments

- Do you have any other comments?

B TOOL INSTRUCTIONS

Thematic Analysis Coding Assistant Tool

The tool trains a machine learning classifier on user-coded sentences in a transcript to code additional sentences you might have missed during thematic analysis. Initially, the model might not be very accurate, but you can keep refining data by re-labelling sentences and re-training the classifier to attempt to improve accuracy. However, please note that the focus of the study is your experience with the reclassification process and your interactions with the model, rather than the accuracy of the classifier.

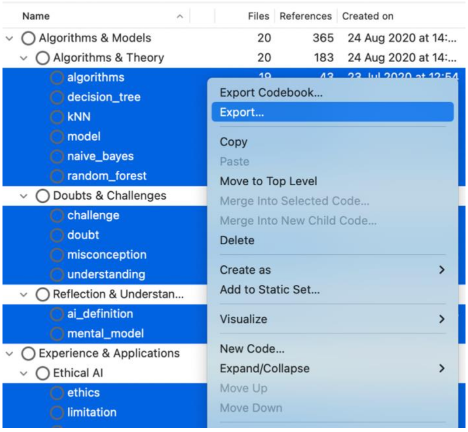
Setup

Step 1: Organise your files

If you used **Microsoft Word** to code your transcript, codes should appear in comments, and the same delimiter should be used to separate multiple codes in the same comment, e.g. "; ".

If you used **NVivo** to code your transcript:

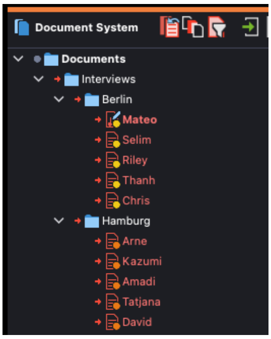
1. Inside NVivo, select all the codes at the lowest level -> right click -> Export…
   (to quickly select all the codes: Ctrl/⌘ + A -> Ctrl/⌘ + click to deselect higher level codes)



2. On Windows, after you click on "Export…", select "Reference View", the "Name" checkbox, and "Folder and Hierarchical Name" from the dropdown list.
3. Export all the codes in a separate folder. The folder should only contain .docx files, one for each code.

If you used **MAXQDA** to code your transcript:

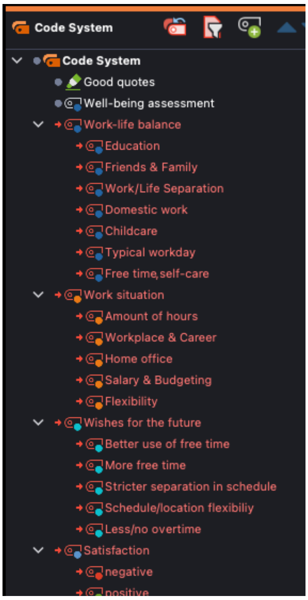
1. Inside MAXQDA, select the transcripts from the Document System pane:



Document System

- Documents
  - Interviews
    - Berlin
      - Mateo
      - Selim
      - Riley
      - Thanh
      - Chris
    - Hamburg
      - Arne
      - Kazumi
      - Amadi
      - Tatjana
      - David

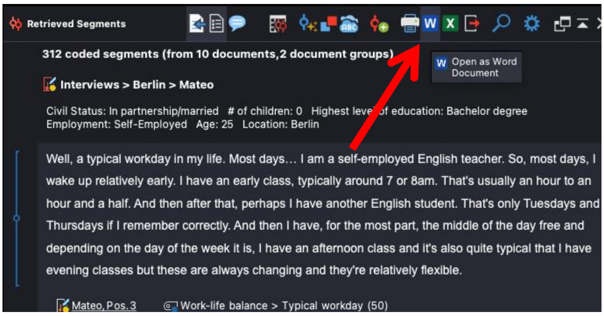
2. Select the codes from the Code System pane:



Code System

- Code System
  - Good quotes
  - Well-being assessment
  - Work-life balance
    - Education
    - Friends & Family
    - Work/Life Separation
    - Domestic work
    - Childcare
    - Typical workday
    - Free time,self-care
  - Work situation
    - Amount of hours
    - Workplace & Career
    - Home office
    - Salary & Budgeting
    - Flexibility
  - Wishes for the future
    - Better use of free time
    - More free time
    - Stricter separation in schedule
    - Schedule/location flexibiliy
    - Less/no overtime
  - Satisfaction
    - negative
    - positive

3. Open the Retreived Segments pane and click on the **W** button to export all segments in a .docx file.



312 coded segments (from 10 documents,2 document groups)

Interviews > Berlin > Mateo

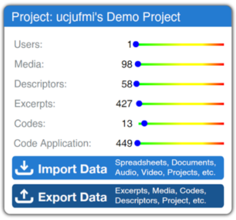
Civil Status: In partnership/married # of children: 0 Highest level of education: Bachelor degree
Employment: Self-Employed Age: 25 Location: Berlin

Well, a typical workday in my life. Most days… I am a self-employed English teacher. So, most days, I wake up relatively early. I have an early class, typically around 7 or 8am. That's usually an hour to an hour and a half. And then after that, perhaps I have another English student. That's only Tuesdays and Thursdays if I remember correctly. And then I have, for the most part, the middle of the day free and depending on the day of the week it is, I have an afternoon class and it's also quite typical that I have evening classes but these are always changing and they're relatively flexible.

Mateo, Pos. 3 Work-life balance > Typical workday (50)

If you used **Dedoose** to code your transcript:

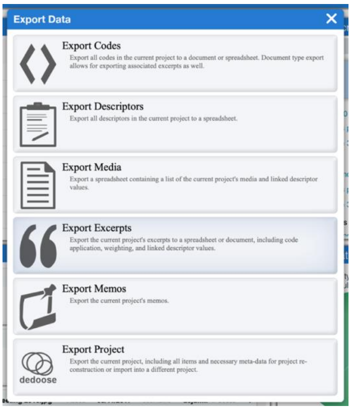
1. Inside Dedoose, click Export Data in the Project pane:


Project: ucjufmi's Demo Project

Users:	1
Media:	98
Descriptors:	58
Excerpts:	427
Codes:	13
Code Application:	449

Import Data
Spreadsheets, Documents, Audio, Video, Projects, etc.
Export Data
Excerpts, Media, Codes, Descriptors, Project, etc.

2. Click Export Excerpts in the popup:


Export Data

Export Codes
Export all codes in the current project to a document or spreadsheet. Document type export allows for exporting associated excerpts as well.

Export Descriptors
Export all descriptors in the current project to a spreadsheet.

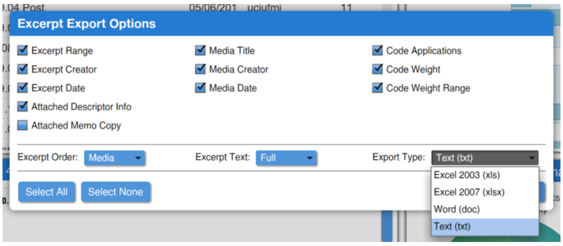
Export Media
Export a spreadsheet containing a list of the current project's media and linked descriptor values.

Export Excerpts
Export the current project's excerpts to a spreadsheet or document, including code application, weighting, and linked descriptor values.

Export Memos
Export the current project's memos.

Export Project
Export the current project, including all items and necessary meta-data for project re-construction or import into a different project.

3. Select Text (txt) under Export Type and click Export to export the excerpts in a .txt file (leave all the checkboxes untouched):



Excerpt Export Options

☑ Excerpt Range
☑ Excerpt Creator
☑ Excerpt Date
☑ Attached Descriptor Info
☐ Attached Memo Copy
☑ Media Title
☑ Media Creator
☑ Media Date
☑ Code Applications
☑ Code Weight
☑ Code Weight Range

Excerpt Order: Media
Excerpt Text: Full
Export Type: Text (txt)

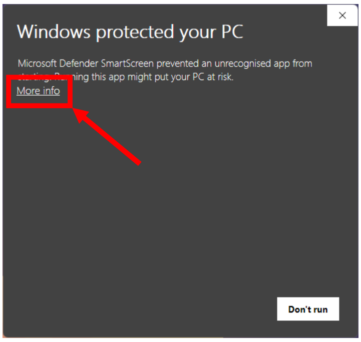
Step 2: Run the tool

Please note that the tool is at an early stage so you might encounter some bugs. If this happens, please send us the error report shown in the error popup, making sure the text does not contain any sensitive data, such as extracts from your transcript.

To install the tool,

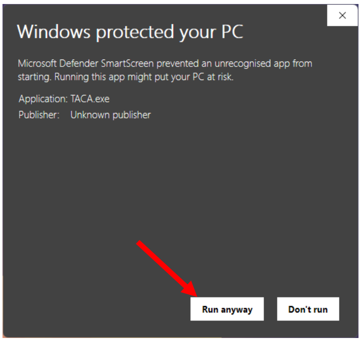
on Windows:

1. Extract TACA.zip in a desired location
2. Navigate inside the TACA directory and run TACA.exe
3. Allow the executable to run:



Windows protected your PC

Microsoft Defender SmartScreen prevented an unrecognised app from starting. Running this app might put your PC at risk.



Windows protected your PC

Microsoft Defender SmartScreen prevented an unrecognised app from starting. Running this app might put your PC at risk.

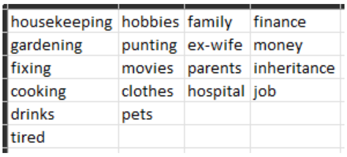
Application: TACA.exe
Publisher: Unknown publisher

on MacOS:

1. Open TACA.dmg
2. Drag the .app into the Applications directory
3. Navigate to Applications and run TACA.app
4. Allow the application to run

Please be patient while the tool loads for the first time. This can take several minutes, and the window might appear blank. When the tool has finished loading, you should see the initial page where you can import your files. In order:

1. Import your transcript .docx file
2. Select whether the transcript was coded using Word, NVivo, MAXQDA or Dedoose
3. If you selected Word, insert the delimiter you used to separate multiple codes in the same comment
4. If you selected NVivo, import the folder containing the codes .docx files
5. If you selected MAXQDA, import the Coded Segments .docx file
6. If you selected Dedoose, import the Excerpts .txt file
7. Edit the codes.csv file automatically generated so that each column header is the name of a theme, and the respective codes appear below, e.g.:



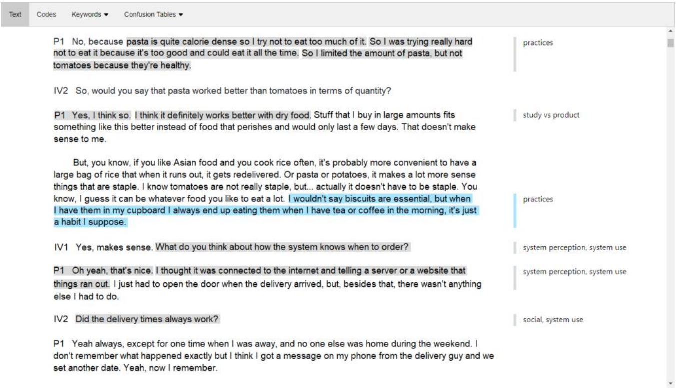
housekeeping	hobbies	family	finance
gardening	punting	ex-wife	money
fixing	movies	parents	inheritance
cooking	clothes	hospital	job
drinks	pets		
tired			

8. Enter meaningless words or terms to be ignored by the machine learning model. These should be separated by a semicolon, e.g. “Interviewer 1;Participant 1”. You can skip this step.
9. Press “Done” to train the model
10. Wait until the model is done training (this can take a while depending on the length of your transcript)

Using the tool

For a video tutorial of the tool, please click here.

Text



P1 No, because pasta is quite calorie dense so I try not to eat too much of it. So I was trying really hard not to eat it because it's too good and could eat it all the time. So I limited the amount of pasta, but not tomatoes because they're healthy. — practices

IV2 So, would you say that pasta worked better than tomatoes in terms of quantity?

P1 Yes, I think so. I think it definitely works better with dry food. Stuff that I buy in large amounts fits something like this better instead of food that perishes and would only last a few days. That doesn't make sense to me. — study vs product

But, you know, if you like Asian food and you cook rice often, it's probably more convenient to have a large bag of rice that when it runs out, it gets redelivered. Or pasta or potatoes, it makes a lot more sense things that are staple. I know tomatoes are not really staple, but... actually it doesn't have to be staple. You know, I guess it can be whatever food you like to eat a lot. I wouldn't say biscuits are essential, but when I have them in my cupboard I always end up eating them when I have tea or coffee in the morning, it's just a habit I suppose. — practices

IV1 Yes, makes sense. What do you think about how the system knows when to order? — system perception, system use

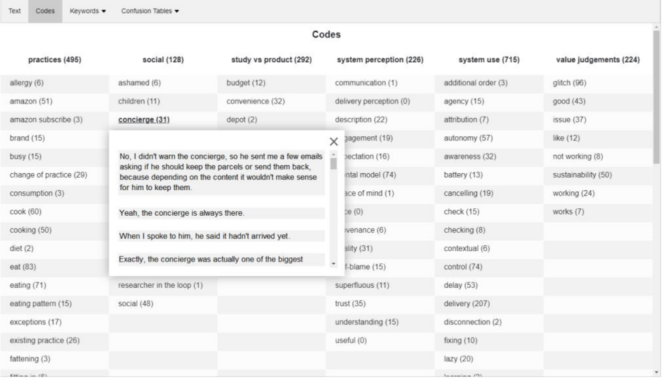
P1 Oh yeah, that's nice. I thought it was connected to the internet and telling a server or a website that things ran out. I just had to open the door when the delivery arrived, but, besides that, there wasn't anything else I had to do. — system perception, system use

IV2 Did the delivery times always work? — social, system use

P1 Yeah always, except for one time when I was away, and no one else was home during the weekend. I don't remember what happened exactly but I think I got a message on my phone from the delivery guy and we set another date. Yeah, now I remember.

The Text tab contains the entire transcript. Sentences you coded manually are highlighted in **grey**, while those coded by the model appear in **blue**. Theme names appear in line with the respective sentences and are also shown in a tooltip on mouseover. The tool works with themes instead of codes to simplify the implementation of the learning algorithm.

Codes



Codes

practices (495)	social (128)	study vs product (292)	system perception (226)	system use (715)	value judgements (224)
allergy (6)	ashamed (6)	budget (12)	communication (1)	additional order (3)	glitch (96)
amazon (51)	children (11)	convenience (32)	delivery perception (0)	agency (15)	good (43)
amazon subscribe (3)	concierge (31)	depot (2)	description (22)	attribution (7)	issue (37)
brand (15)			gagement (19)	autonomy (57)	like (12)
busy (15)			pectation (16)	awareness (32)	not working (8)
change of practice (29)			ntal model (74)	battery (13)	sustainability (50)
consumption (3)			ace of mind (1)	cancelling (19)	working (24)
cook (60)			ce (0)	check (15)	works (7)
cooking (50)			venance (6)	checking (8)	
diet (2)			ality (31)	contextual (6)	
eat (83)			f-blame (15)	control (74)	
eating (71)	researcher in the loop (1)		superfluous (11)	delay (53)	
eating pattern (15)	social (48)		trust (35)	delivery (207)	
exceptions (17)			understanding (15)	disconnection (2)	
existing practice (26)			useful (0)	fixing (10)	
fattening (3)				lazy (20)	
fitting in (6)				learning (2)	

No, I didn't warn the concierge, so he sent me a few emails asking if he should keep the parcels or send them back, because depending on the content it wouldn't make sense for him to keep them.

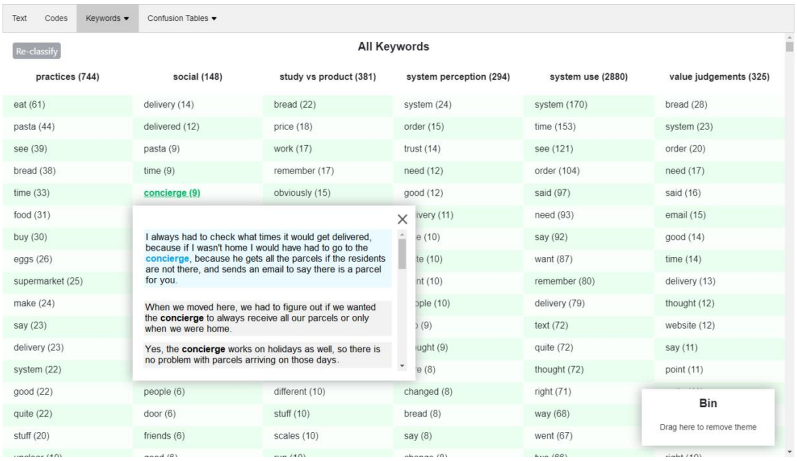
Yeah, the concierge is always there.

When I spoke to him, he said it hadn't arrived yet.

Exactly, the concierge was actually one of the biggest

The codes tab contains the table of codes you have imported. Each theme and code show a counter indicating the number of sentences (this counter can be 0 if no sentences were found with that code). You can click a code to reveal the sentences you manually coded in a tooltip.

Keywords Tables



All Keywords

practices (744)	social (148)	study vs product (381)	system perception (294)	system use (2880)	value judgements (325)
eat (61)	delivery (14)	bread (22)	system (24)	system (170)	bread (28)
pasta (44)	delivered (12)	price (18)	order (15)	time (153)	system (23)
see (39)	pasta (9)	work (17)	trust (14)	see (121)	order (20)
bread (38)	time (9)	remember (17)	need (12)	order (104)	need (17)
time (33)	concierge (9)	obviously (15)	good (12)	said (97)	said (16)
food (31)			ivery (11)	need (93)	email (15)
buy (30)			e (10)	say (92)	good (14)
eggs (26)			te (10)	want (87)	time (14)
supermarket (25)			nt (10)	remember (80)	delivery (13)
make (24)			ople (10)	delivery (79)	thought (12)
say (23)			(9)	text (72)	website (12)
delivery (23)			ught (9)	quite (72)	say (11)
system (22)			e (8)	thought (72)	point (11)
good (22)	people (6)	different (10)	changed (8)	right (71)	
quite (22)	door (6)	stuff (10)	bread (8)	way (68)	
stuff (20)	friends (6)	scales (10)	say (8)	went (67)	
unclear (10)	good (6)	run (10)	change (8)	two (66)	right (10)

I always had to check what times it would get delivered, because if I wasn't home I would have had to go to the **concierge**, because he gets all the parcels if the residents are not there, and sends an email to say there is a parcel for you.

When we moved here, we had to figure out if we wanted the **concierge** to always receive all our parcels or only when we were home.

Yes, the **concierge** works on holidays as well, so there is no problem with parcels arriving on those days.

The Keywords tabs contain three tables:

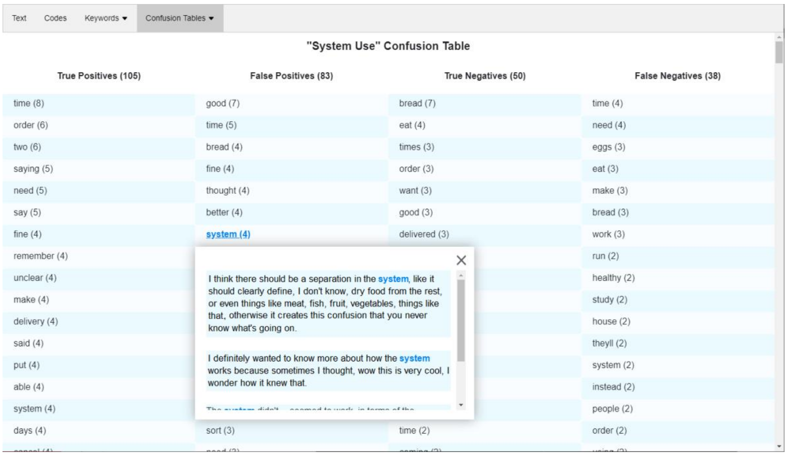
1. "**Predict Keywords**" containing only sentences coded by the model
2. "**Train Keywords**" containing only sentences you manually coded
3. "**All Keywords**" containing both types of sentences

Here, the most frequent words are shown under each theme, along with a counter indicating the number of sentences that contain them. You can click on a word to reveal these sentences in a tooltip. Sentences coded by the model have a **blue** background, while those manually coded have a **grey** background.

You can re-label sentences to different themes either by dragging and dropping single sentences from the tooltip to a different column/bin, or dragging and dropping keywords from one column to another, or to the bin. Moving sentences to the bin removes the theme from those sentences. Moving a keyword is equal to moving the entire list of sentences that contain it. You might see several meaningless keywords that you might have forgotten to include in the keywords filter. Please ignore these and focus on the meaningful keywords/sentences you would like to re-label.

After you are finished re-labelling sentences, click the "Re-classify" button to re-train the machine learning model. The tool will update all the tabs once it is finished loading. Significant changes in the Keywords tables after re-training will be shown with highlighted table cells.

Confusion Tables



"System Use" Confusion Table

True Positives (105)	False Positives (83)	True Negatives (50)	False Negatives (38)
time (8)	good (7)	bread (7)	time (4)
order (6)	time (5)	eat (4)	need (4)
two (6)	bread (4)	times (3)	eggs (3)
saying (5)	fine (4)	order (3)	eat (3)
need (5)	thought (4)	want (3)	make (3)
say (5)	better (4)	good (3)	bread (3)
fine (4)	system (4)	delivered (3)	work (3)
remember (4)			run (2)
unclear (4)			healthy (2)
make (4)			study (2)
delivery (4)			house (2)
said (4)			theyll (2)
put (4)			system (2)
able (4)			instead (2)
system (4)			people (2)
days (4)	sort (3)	time (2)	order (2)

I think there should be a separation in the system, like it should clearly define, I don't know, dry food from the rest, or even things like meat, fish, fruit, vegetables, things like that, otherwise it creates this confusion that you never know what's going on.

I definitely wanted to know more about how the system works because sometimes I thought, wow this is very cool, I wonder how it knew that.

The Confusion Tables tabs contain a table version of confusion matrices for each theme. Confusion matrices are a way to evaluate the performance of the classifier. The model takes 20% of the sentences you manually coded ignoring the codes, and tries to guess them itself to see what it gets right. Confusion matrices are made of 4 quadrants, in this case columns:

- True Positives: Sentences the model **should have coded** in this theme, and **did**
- False Positives: Sentences the model **should not have coded** in this theme, but **did**
- True Negatives: Sentences the model **should not have coded** in this theme, and **did not**
- False Negatives: Sentences the model **should have coded** in this theme, but **did not**

Words are sorted by frequency in each column, along with a counter. You can click on each word to reveal the sentences that contain it.

Received July 2024; revised October 2024; accepted December 2024