

Evaluating Interaction with Machine Learning Text Classifiers and Interpretability Techniques

Federico Milana

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London.

UCL Interaction Centre
University College London

February 24, 2025

I, Federico Milana, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

As Machine Learning (ML) becomes increasingly integrated into society and more users interact with ML-driven systems, understanding how they perceive and engage with these technologies grows increasingly important. This Ph.D. thesis explores user experience, usability, interpretability, and cognitive biases of ML text classifiers through two research projects based on a desktop application developed to support thematic analysis, and a separate user evaluation of interpretability techniques.

The Thematic Analysis Coding Assistant (TACA) enables users to import an initial thematic analysis, extracts labelled sentences, trains an offline gradient boosting classifier, and generates coding suggestions. Users can iteratively re-train the model after re-labelling individual or batches of sentences. A user study run with 20 non-ML expert participants revealed that participants critically reflected on their analysis, gained new thematic insights, and adapted their interpretative stance, while also showing misconceptions about ML concepts, positivist views, and self-blame for poor model performance.

A second study reports on an autoethnography of the use of TACA, revealing different re-labeling and model inspection strategies, reflecting on potential structural changes to the analysis, and examining the positionality of the user as a developer, a researcher, and a participant. The findings provide complementary insights into how ML can support and challenge analytical processes, personal reflections, and perceptions of the model.

Building on the findings of the first two studies, a third study evaluates two popular local interpretability methods in text classification, LIME and SHAP, and

a proposed global method using LLM-generated summaries based on LIME importance weights. Among 128 participants, those without explanations identified broader topics, while those using LIME and SHAP focused on individual terms, and those using summaries identified more features overall. However, none of the methods significantly improved model prediction accuracy.

Together, these studies contribute to seminal research on understanding the perception and interaction with ML and the implications of system and interaction design to improve the understanding of ML concepts.

Impact Statement

Recent advances in ML algorithms, computational power, and data availability have driven what is now considered an ongoing AI spring. ML is increasingly used in more aspects of society and is transforming industries, ranging from healthcare and finance to education and entertainment, reshaping how decisions are made and services are delivered.

As the number of applications for ML continues to grow, so does the number of users exposed to these systems. Although the underlying theory behind ML is highly advanced, users are typically not required to understand the technical details to interact even with extremely large and complex models. This is significant considering the particularly high stakes in critical application areas where some of these models are used, such as medical diagnosis, legal decision-making, and autonomous driving. Consequently, understanding how users perceive and engage with ML is arguably just as important as advances and breakthroughs in AI itself.

This thesis contributes to existing research on human-AI interaction by specifically evaluating user experience with ML text classifiers. ML in text processing has experienced remarkable success in recent years, revolutionising natural language understanding, machine translation, sentiment analysis, and automated content generation. Text classification, in particular, plays a crucial role in many applications ranging from categorizing large volumes of unstructured text to supporting automated decision-making systems in various domains involving user interaction, including spam detection, customer support automation, and content moderation.

The contributions of this thesis are manifold and will potentially gain more significance over time as ML continues to be used more frequently and in more ap-

plication areas. It is hoped that the different quantitative, qualitative and self-study approaches used in this work will inspire the research community to take advantage of the wide range of research methods available in HCI to evaluate interaction with AI. All the design implications derived from the findings are provided in hope of guiding future research and development of AI systems that prioritise user-centred design, improve transparency, and facilitate the integration of new technologies in modern society.

The impact of this work extends beyond academia by informing designers and developers in various sectors of the industry on how to create more intuitive, transparent, and effective applications of AI that better align with user needs and expectations. Companies developing these systems can benefit from the findings by understanding how end users, from professionals to everyday consumers, perceive and interact with ML, ultimately enhancing user satisfaction and promoting broader adoption of AI technologies.

Acknowledgements

I would like to thank my supervisors Enrico Costanza, Mirco Musolesi, and Amid Ayobi for their efforts throughout the course of my Ph.D.. I truly could not have asked for better guidance, support, and inspiration for my work. The diversity and depth of their knowledge and skills are the main reason for my growth as a researcher and I feel sincerely lucky to have worked with them.

I would like to thank my parents and my brother for their many words of encouragement and their constant belief in me. Their presence gave me the strength and motivation to persevere through every challenge.

I would also like to thank all my friends for their invaluable ability to lift my spirits and celebrate my achievements along the way.

Lastly, I would like to express my gratitude for the financial support provided by the Engineering and Physical Sciences Research Council (EPSRC).

Contents

1	Introduction	16
1.1	Research Questions	19
1.2	Thesis Structure	22
1.3	Research Contributions	23
2	Literature Review	26
2.1	Background	26
2.2	Interactive Machine Learning	29
2.2.1	System Design	30
2.2.2	Model Inspection	31
2.2.3	Text Applications	33
2.3	Explainable Artificial Intelligence	35
2.3.1	Interpretability in Text Classification	37
2.3.2	User Studies	38
2.4	Summary and Discussion	40
3	Understanding Interaction with Machine Learning through a Thematic Analysis Coding Assistant: A User Study	43
3.1	Qualitative Data Analysis as an Application Area for Interactive Machine Learning	45
3.2	The Thematic Analysis Coding Assistant (TACA)	46
3.2.1	Implementation Details	49
3.3	Study	52

	<i>Contents</i>	9
3.3.1	Participants	52
3.3.2	Procedure	52
3.3.3	Analysis	55
3.4	Findings	55
3.4.1	System Usage from Automatic Interaction Logs	55
3.4.2	Evaluation Strategies for Model Inspection and Reflections on Machine Learning Output	57
3.4.3	Benefits and Challenges of Data Aggregation	59
3.4.4	Perception of the Machine Learning Model	60
3.4.5	Personal Blame for Poor Model Performance	61
3.4.6	Perceived and Anticipated Use of Interactive Machine Learning in Qualitative Data Analysis	62
3.5	Discussion	63
3.5.1	Supporting Reflexivity with Interactive Machine Learning .	63
3.5.2	Balancing Objectivity and Subjectivity in Interactive Ma- chine Learning	65
3.5.3	Understanding Perceptions of Machine Learning through User Interface Features	67
3.6	Conclusion	70
4	An Autoethnography on the Thematic Analysis Coding Assistant	72
4.1	Self-study Research Methods	73
4.1.1	Autoethnography	73
4.1.2	Self-study in Human-Computer Interaction	74
4.1.3	Self-built systems	75
4.1.4	Self-study in Artificial Intelligence	76
4.2	Motivation for Autoethnography	77
4.3	Reflexivity and Positionality	79
4.4	Method	81
4.4.1	Previous Qualitative Data Analysis Results	81
4.4.2	Autoethnography Data Collection	83

	<i>Contents</i>	10
4.5 Findings	86	
4.5.1 System Usage	86	
4.5.2 Interacting with TACA	87	
4.5.3 Perception of the Machine Learning Model	93	
4.5.4 Emerging Insights and Analytical Reflections	97	
4.6 Discussion	101	
4.6.1 Reflecting on the Different Roles in the Autoethnography .	101	
4.6.2 TACA Effectively Supports the Refinement of Qualitative Data Analysis	104	
4.6.3 Implications of Design Choices on Interactive Machine Learning	107	
4.6.4 Machine Learning and Human Interpretation	110	
4.7 Conclusion	113	
5 Evaluating Model-Agnostic Interpretability Techniques for Machine Learning Text Classification: A User Study on Predicting Model Outcome		115
5.1 Method	117	
5.1.1 Materials	118	
5.1.2 Model Architecture and Training	119	
5.1.3 Tasks	122	
5.1.4 Conditions	124	
5.1.5 Participants	125	
5.1.6 Procedure	125	
5.1.7 Pilot Studies	126	
5.2 Results	128	
5.2.1 Outcome Prediction Accuracy	128	
5.2.2 Confidence in Predictions	130	
5.2.3 Perceived Usefulness of Example Categories and Word Importance Heat Maps	131	
5.2.4 Mentioned Features	131	

<i>Contents</i>	<i>11</i>
5.3 Discussion	134
5.3.1 No Effects of Techniques on Model Outcome Prediction . .	135
5.3.2 Effects of Techniques on Feature Attention	139
5.4 Conclusion	141
6 General Discussion and Conclusions	143
6.1 Summary and Key Findings	143
6.2 Design Implications	146
6.2.1 Chapter 3 Design Implications	146
6.2.2 Chapter 4 Design Implications	148
6.2.3 Chapter 5 Design Implications	149
6.3 Discussion	149
6.3.1 Extended Benefits of AI Beyond Model Performance . . .	149
6.3.2 Challenges in Perceiving AI as an Objective Authority . .	151
6.3.3 Model Transparency to Manage User Perceptions of ML .	153
6.4 Limitations	156
6.5 Future Work	158
6.6 Conclusion	161
Appendices	163
A Chapter 3 Supplementary Material	163
A.1 User Study Participant Information Sheet (Participants' Own Data) .	164
A.2 User Study Participant Information Sheet (Newspaper Restaurant Reviews)	166
A.3 User Study Informed Consent Form	168
A.4 Tool Instructions	169
A.5 Semi-structured Interview Script	177
B Chapter 4 Supplementary Material	182
B.1 Manual Qualitative Data Analysis Themes and Codes	183

<i>Contents</i>	12
C Chapter 5 Supplementary Material	184
C.1 User Study Participant Information Sheet	185
C.2 User Study Informed Consent Form	187
C.3 User Study Instructions (Heat Maps)	189
C.4 User Study Instructions (Summaries)	191
C.5 LLM-Generated Summaries	193
C.5.1 Class: Opinions	193
C.5.2 Class: Place	194
C.5.3 Prompt Used	195
C.6 Example User Study Page (LIME)	196
C.7 Prediction Accuracy Across Different Interpretability Techniques . .	200
C.8 Pilot Studies Results	201
C.8.1 Pilot Study 1	201
C.8.2 Pilot Study 2	203
Bibliography	205

List of Figures

3.1	Text page showing highlighted user-coded sentences in grey and classified sentences in blue.	47
3.2	Codes page showing a lookup table for user-defined codes and respective themes.	48
3.3	All Keywords Table page showing the most frequently occurring terms for each theme.	49
3.4	Confusion Table page showing the most frequently occurring terms for each confusion matrix quadrant of the selected theme.	50
3.5	Average number of data points re-labelled in each table.	57
4.1	Screenshot of the Codes Table	88
4.2	Example of a code including a sentence taken out of context: “And got the washing machine over there”.	89
4.3	Screenshot of the Train Keywords Table	90
4.4	Screenshot of the Train Keywords Table after the initial re-labelling process.	91
4.5	Screenshot of the Confusion Table for the theme “Interacting with the Model”.	93
4.6	Screenshot of the Predict Keywords Table after the initial re-labelling and re-training process.	95
5.1	Word importance heat maps generated from LIME and SHAP weights according to the class “opinions”.	121
5.2	Distributions of total correct predictions to all questions by condition.	129

5.3	Distributions of average confidence scores (1-4) to all questions by condition.	130
5.4	The average frequency of terms, topics and language feature types by condition.	133
5.5	Normalised frequencies of salient features mentioned by participants for samples about “opinions” (top 2) and “place” (bottom 2). Left: specific terms, right: topics and language.	134
C.1	Distributions of total correct predictions to all questions by condition in the first pilot study.	202
C.2	Distributions of total correct predictions to all questions by condition in the second pilot study.	204

List of Tables

3.1	User study participants information.	53
5.1	Topics manually labelled belonging to class labels “opinions” and “place”.	119
6.1	A summarised list of contributions made by this thesis.	147
B.1	Themes (header) and codes identified in the manual thematic anal- ysis performed on the interview transcripts in the user study.	183
C.1	Comparison of prediction accuracy based on different interpretabil- ity techniques for all tasks.	200
C.2	Comparison of prediction accuracy based on different interpretabil- ity techniques for all tasks in the first pilot study.	201
C.3	ANOVA results for the effect of condition on the number of correct predictions in the first pilot study. Note: Type III Sum of Squares.	202
C.4	Comparison of prediction accuracy based on different interpretabil- ity techniques for all tasks in the second pilot study.	203
C.5	ANOVA results for the effect of condition on correct predictions in the second pilot study. Note: Type III Sum of Squares.	204

Chapter 1

Introduction

Recent advances in algorithms, computational power, and data availability have made Machine Learning (ML) ubiquitous in the digital landscape of today, with applications that span an ever-growing number of domains and industries. The underlying theory of ML is highly advanced and requires familiarity with mathematical concepts, including linear algebra, calculus, and probability, that even many software engineers may not possess. However, users are now increasingly exposed to systems that do not require theoretical knowledge to interact and engage with. This user base, spanning various industries and everyday applications, often relies on AI systems to make critical decisions without fully understanding how they function.

The growing interaction between humans and Artificial Intelligence (AI) brings new challenges in ensuring that users can meaningfully engage with, trust, and interpret the output of ML systems. This is especially true as models continue to grow in size and complexity, as recently demonstrated in Large Language Models (LLMs), and the high-stake scenarios that they are employed in, such as medical diagnosis, legal decision-making, and autonomous driving. Understanding human-AI interaction becomes crucial to design for effective collaboration by bridging the gap between complex algorithms and user comprehension, to the point where it is arguably just as important as the technical advances and breakthroughs in AI. This thesis focuses on two aspects of ML that provide opportunities for deeper user engagement and understanding: Interactive ML (IML) and Explainable Artificial Intelligence (XAI).

Currently, most ML applications are based on models trained on large data sets, which seem to keep growing in size. The dependence of model performance on the size of the training set is widely identified as one of the limitations of ML (Mastorakis, 2018). As a response, there is growing interest in achieving high performance by customising models trained on smaller data sets (Y. Zhang and Ling, 2018; Phung and Rhee, 2019; Choi and Ma, 2020; C. Y. Liao, P. Liu, and Wu, 2020). IML 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 (Amershi, Cakmak, et al., 2014). IML involves end-users in an iterative and incremental learning process and leverages human feedback to drive ML. 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. In addition to creative applications, where user customisation of ML models is particularly beneficial (Caramiaux and Tanaka, 2013), the “human-in-the-loop” approach is showing promising results in health informatics (Holzinger, 2016) and environmental sciences (Medeiros et al., 2020), where only small data sets are available and problems are characterised 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 (Ware et al., 2001).

Compared to earlier implementations of ML models, such as Naive Bayes classifiers, k-Nearest Neighbors and Decision Trees, the model architectures used today are significantly more sophisticated. Although increasing algorithmic complexity has achieved remarkable results in ML applications, it has also made it increasingly difficult to understand, trust, and explain the output of these models. As the inner workings remain largely inaccessible, sometimes even for experts, interpretability remains a fundamental open challenge in ML (Lipton, 2018). Interpretability is crucial because it directly impacts trust, reliability, robustness, causality, and usability of models (Doshi-Velez and B. Kim, 2017), particularly in high-stakes areas (Samek, Wiegand, and Müller, 2017). When models are interpretable, stakeholders,

including practitioners, regulators, and end users can understand how decisions are made, which is essential to validate the correctness of the model and ensure that it operates fairly and ethically (Langer et al., 2021).

This thesis narrows its focus specifically to one of the most successful and widely used applications of ML: text classification (Guidotti et al., 2018). In recent years, ML in text processing has achieved remarkable success, transforming the area of Natural Language Processing (NLP). Sentiment analysis, spam detection, topic categorisation, language detection, and social media moderation are just a few examples of how text classifiers are used today to improve customer experiences, streamline operations, or gain actionable insights from vast amounts of unstructured data.

The data sets used for ML classification in prior research typically have a well-defined ground truth, for example news articles labelled by topic or reviews that are clearly positive or negative. However, in reality, many other data sets are usually more complex, and numerous external factors can significantly influence the process of annotation. Factors such as cultural context, linguistic nuances, or personal interpretation by annotators can introduce ambiguity and subjectivity in the manual labeling of training data sets (Miceli, Schuessler, and T. Yang, 2020). Additionally, data used for AI tools in creative industries is often inherently noisy, non-stationary, or incomplete (Caramiaux, Lotte, et al., 2019). To better reflect real-world scenarios where the ground truth is not always well-defined, provide more realistic assessments of user behaviour, and yield more actionable results, the studies reported in this thesis intentionally make use of data sets that are ambiguous and subject to interpretation.

The design of user studies around AI systems has been recognised as a challenge (Kittley-Davies et al., 2019), as it is important to support positive experiences of participants in the pursuit of ecological validity. This thesis focuses on Qualitative Data Analysis (QDA) as a domain for the studies. Not only has previous work highlighted the potential benefits of applying ML to QDA (Chen et al., 2018; Gebreegziabher et al., 2023), but QDA also is particularly suitable in capturing the

individuality of the user by enabling interaction with ML models according to individual perspectives and interpretation of ambiguous data. However, although QDA serves as a valuable application domain for the studies, it should be emphasised that the main interest lies in the interaction with ML.

1.1 Research Questions

The primary aim of this thesis is to evaluate interaction with ML in terms of how users perceive, engage with, and interpret ML text classification models. The following research questions (RQs) were formulated over the course of the Ph.D., with the later questions shaped and refined by the findings of the earlier studies. The first three RQs were investigated through the Thematic Analysis Coding Assistant (TACA), a fully functioning IML GUI desktop application developed to assist the coding phase of the analysis by training a text classifier on an initial coding phase to provide additional coding suggestions. The four questions are addressed in the context of QDA as an application for text classification.

RQ1: How do non-expert users perceive ML when analysing ambiguous data?

Where by “perceive” includes how non-expert users understand the contribution of ML to their analytical process. For example, whether they consider ML an analytical tool, a supplementary assistant, a collaborative partner, or an authoritative source of insights. Shaped by calls from the literature to enhance, rather than supplant, the work of human coders (Lewis, Zamith, and Hermida, 2013), this RQ was meant to validate whether users recognise the subjectivity inherent the data, or if they view the output of the model as objective, unbiased contributions. This RQ leads into the following, consequent question:

RQ2: How do non-expert users’ perceptions of ML influence their interaction with it?

This RQ explores the consequences of non-expert perceptions of the ML model, including biases and expectations, specifically how they influence how they interpret the output of the ML model and consequently interact with it through the

feedback assignment phase of the IML cycle. The following RQ was formulated to gain complementary insights into IML:

RQ3: How can IML be used to support the analysis of ambiguous data?

This RQ explores whether IML can lead to new thematic insights and a more critical evaluation of subjective and ambiguous data, enabling deeper, more nuanced interpretations that may not emerge from manual methods. It highlights distinctions between manual and ML-supported approaches to QDA and questions how the feedback-driven, iterative nature of ML analysis encourages critical thinking and alternative strategies. The findings revealed from this RQ raised broader questions about the effects of ML on analytical workflows and outcomes, such as whether the use of ML introduces new ways of thinking about the data by challenging existing assumptions and encouraging more reflexive approaches.

Building on the results of the first three RQs, the final question extends the focus onto XAI, exploring how interpretability could support IML for QDA:

RQ4: How do interpretability techniques affect users' ability to predict ML model behaviour?

The formulation of this RQ was mainly influenced by previous work on Layer-Wise Relevance Propagation (LRP) saliency maps used in Convolutional Neural Network (CNN) image classification (Alqaraawi et al., 2020). Many user studies in the literature on XAI measure user understanding of explanations (Lage et al., 2019; Poursabzi-Sangdeh, Daniel G. Goldstein, et al., 2021b) or self-reported trust (Nourani et al., 2019; Papenmeier, Englebienne, and Seifert, 2019) as a proxy for usefulness of interpretability techniques. However, recent work has recently revealed significant limitations of traditional methods for evaluating model explanations involving subjective feedback (Lakkaraju and Bastani, 2020; Schneider, Meske, and Vlachos, 2021). This RQ is based on a previously proposed metric to assess the interpretability of a system according to which, if users truly understand how the system functions, they should be able to accurately predict its output (Muramatsu and Pratt, 2001). This method, known as “forward simulation”, involves asking users to anticipate or simulate the model’s output, thereby providing an ob-

jective measure of interpretability (Belle and Papantonis, 2021). Although many recent studies have used forward simulation to evaluate interpretability in various domains (B. Kim, Khanna, and Koyejo, 2016; Poursabzi-Sangdeh, Daniel G Goldstein, et al., 2021a; Buçinca, Lin, et al., 2020; Alqaraawi et al., 2020; Waa et al., 2021), it has yet to be applied in the context of interpretability for text classification.

The RQ was answered by addressing the following sub-questions:

- RQ4.1: Do SHAP and LIME generated word importance heat maps assist participants in predicting the outcome of a text classifier?
- RQ4.2: Are LLM-generated summaries of LIME word importance weights an effective interpretability technique?
- RQ4.3: What are the effects of interpretability techniques on the confidence of predictions of the model outcome?
- RQ4.4: How do different interpretability techniques influence users' attention towards specific features, and what effect does this have on their ability to understand overall model behaviour?

RQ4.1 was formulated around LIME and SHAP as these are the two most popular local interpretability techniques today (Salih et al., 2024; Aechtner et al., 2022; Cesarini et al., 2024). Given that the limitations of local explanations have been widely discussed in the literature (Ribeiro, Singh, and Guestrin, 2018; Alqaraawi et al., 2020; Chromik et al., 2021), RQ4.2 leverages recent advancements in LLMs to evaluate a novel global interpretability technique using LLM-generated summaries of words according to their LIME importance weights from the training data set. RQ4.3 and RQ4.4 examine the impact of interpretability techniques on users' confidence in their predictions and how these techniques influence their focus on specific features, assessing how this attention affects their overall understanding of model behaviour. RQ4.3 was motivated by previous work on misleading explanations which artificially increase user confidence (Schneider, Meske, and Vlachos, 2021), while RQ4.4 was included to explain the results of RQ4.1 and RQ4.2, in a similar

fashion to the study on saliency maps applied to image classification (Alqaraawi et al., 2020).

1.2 Thesis Structure

The thesis follows three main studies to provide a comprehensive understanding of the key questions, findings, and their implications. Each study is presented in its own chapter.

Chapter 2 reviews the relevant literature on human-AI interaction, IML and XAI, examining key concepts, theories, and prior research in these areas. Identifying gaps in the current research, it lays the foundation for the studies presented in the later chapters.

Chapter 3 documents the Study 1, which investigates how non-experts in ML engage with TACA, to address **RQ1** and **RQ2**. This study explores user interactions, behaviour, and perceptions of 20 participants who applied IML to QDA. The chapter outlines the design and implementation of TACA, describing novel UI features that support iterative re-labeling, model inspection, and feedback assignment. With qualitative and quantitative methods, the chapter examines the participants' understanding of IML processes and their experiences evaluating and refining the model. The discussion highlights reflexivity in reassessing the analysis based on contrasting coding suggestions. It also explores the tension between the subjective data and the perceived objectivity of the model, noting some over-reliance on ML outputs. Finally, it considers to what extent UI features shaped engagement, supporting critical evaluation of model performance.

The participants in Study 1 exhibited limited engagement with the IML aspects of TACA as a consequence of their perceptions of the model as an objective and authoritarian source of advice. To address this limitation, among others, **Chapter 4** builds on the previous chapter by documenting Study 2: an autoethnographic account of TACA framed within the roles of developer, researcher and participant. This chapter addresses **RQ3** by reporting on how sustained interaction with TACA led to deeper insights into the qualitative analysis performed in the previous study

and describing how ML can be applied to analytical processes like QDA. The iterative engagement with the system revealed key differences between non-expert user interactions and those of a user with knowledge in ML, and how these influenced re-labeling strategies, decision-making, model perception and evaluation. The chapter also explores how design choices shaped user interaction, emphasising the importance of balancing usability and transparency in IML systems to support reflexivity and data analysis.

Chapter 5 presents Study 3: an online user study evaluating model-agnostic interpretability techniques in the context of text classification. The study addresses **RQ4** by comparing two widely used techniques, LIME and SHAP, and a novel technique using LLM-generated summaries, in predicting model outcome. The chapter discusses the design and methodology of the experiment, in which 128 participants engaged with different explanations. Through a quantitative and qualitative analysis, the results indicate that different explanations increased user awareness of specific features, but they did not significantly improve the accuracy of outcome prediction. The chapter includes a discussion on user interaction with local versus global explanations, and the limitations of current techniques in handling the complexity of textual data.

Chapter 6 concludes the thesis by summarising the key findings and contributions, discussing the implications for the design of AI systems, and suggesting avenues for future research.

1.3 Research Contributions

The work presented in this thesis provides several contributions to the growing body of literature on human-AI interaction. These include a system contribution in TACA, a fully-functional IML application designed and implemented to answer **RQ1-3**. Significant effort was made to package existing ML Python libraries into executables, an uncommon process that extends beyond their typical use case. The contributions also include the use of a wide range of complementary methodological approaches in quantitative-, qualitative- and self-study-based analyses, demon-

ing that a combination of different research methods can provide a more comprehensive understanding of the field. Notably, the autoethnography reported in this thesis is the first on IML. Empirical contributions are reported in Chapters 3-5 and summarised below.

Chapter 3 shows that TACA was effective in exposing the participants to IML and applying it on qualitative data sets. Specific UI features, such as data aggregation through frequency-based keywords, promoted a thorough examination of the data, facilitating the evaluation of the model and the assignment of feedback during the IML cycle. The study highlights important misconceptions regarding the functionality of the model. Significantly, the 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, answering **RQ1** and **RQ2**. These findings were made possible due to the ambiguous ground truth in the data, which was open to interpretation, allowing participants to critically reflect on their own analytical decisions in relation to the output of the model. Based on this understanding, the chapter elaborates on how applications could be designed to improve the understanding of ML concepts and foster reflexive work practices beyond the scope of QDA.

Chapter 4 answers **RQ3** by systematically describing and analysing how the developer of TACA used and experienced the system through sustained and genuine use. Shaped by the roles of developer, researcher and participant, as well as experience in ML, the interactions with the tool differed significantly from those of the participants in every phase of the IML cycle. The iterative engagement with the model prompted reflections on potential structural changes to the analysis and encouraged reflexivity. Unlike the experiences of the participants, this was not the result of viewing the model as an external, objective source of advice, but rather a consequence of reviewing both the suggestions of the model and the data used for training. The design choices made in TACA highlighted the importance of balancing transparency and usability to support both efficiency and accuracy in IML systems. More generally, the chapter argues that ML relies on the patterns in the

training data and therefore only extends human interpretation without challenging existing assumptions. Although ML can support analysis, uncovering radical new insights remains a human skill, especially when interpreting ambiguous data that requires critical thinking and domain expertise.

Chapter 5 answers **RQ4** by reporting on a thorough evaluation of LIME and SHAP, the interpretability techniques most commonly used in ML text classification, in addition to presenting a novel technique involving LLM-based summarisation. The results suggest that LIME, SHAP, and LLM-generated summaries have a very limited impact in text classification. The explanations guided the participants' attention toward specific features, potentially at the expense of a broader understanding of the behaviour of the model. The large number of features (word embeddings) makes it unlikely that specific features in the LIME and SHAP examples, like individual words, reappear in the prediction tasks. On the other hand, the generated summaries helped participants identify more features overall, but these were only the most prominent, omitting rare but important patterns in the data. The findings confirm the limitations of current interpretability techniques and the need for future research to develop approaches that provide detailed insights while also offering a broader perspective into the inner workings of ML models.

Chapter 3 was presented as a paper to appear in the ACM Conference on Computer Supported Cooperative Work (CSCW) in 2025:

Federico Milana, Enrico Costanza, Mirco Musolesi, and Amid Ayobi.
“Understanding Interaction with ML through a Thematic Analysis Coding Assistant: A User Study”. *Proceedings of the ACM on Human-Computer Interaction (CSCW)*

Chapter 4 has been presented as a paper under review for the ACM Conference on Intelligent User Interfaces (IUI) in 2025:

Federico Milana, Enrico Costanza, Mirco Musolesi, and Amid Ayobi.
“Understanding Interactive Machine Learning through an Autoethnography of the Thematic Analysis Coding Assistant (TACA)”. *Proceedings of the ACM on Human-Computer Interaction (IUI)*

Chapter 2

Literature Review

2.1 Background

AI is becoming a central focus in HCI and UX design to the point where some researchers are now advocating that the priority of current HCI work should be to transition from conventional human interaction with non-AI computing systems to interaction with AI systems (Wei Xu and Gao, 2023).

Human-Centred AI is an emerging discipline that prioritises human needs and values by ensuring AI systems are explainable, comprehensible, useful, and usable. The literature calls for HCI practitioners to proactively contribute to AI research and development to define UX criteria and iteratively test and optimise training data and algorithms. This approach aims to mitigate extreme algorithmic bias while considering privacy, security, environmental protection, social justice, and human rights (Riedl, 2019; Xu, 2019; Shneiderman, 2020).

A key aspect of Human-Centred AI is rethinking ML algorithms and their interfaces based on human goals, contexts, and workflows (Gillies et al., 2016). This perspective prioritises the design of fair, interactive, collaborative, and transparent AI systems that give full agency to stakeholders. As highlighted by Bulathwela et al., 2024, this approach is crucial for fostering equitable and inclusive applications of AI, particularly in domains like education. However, despite the central role humans play in the development, deployment, and use of AI, HCI is often not a core component of AI research (Inkpen et al., 2019). This disconnect can result

in AI systems misaligned with user needs, behaviours, and expectations, leading to reduced usability, trust, and overall effectiveness. To address this, it is necessary to evaluate AI as part of a larger socio-technical system, ensuring that user interactions are integral to system design and refinement.

Human-AI interaction is considered to be uniquely difficult to design. Recent work has categorised challenges as 1) capability uncertainty and 2) output complexity (Q. Yang, Steinfeld, et al., 2020). These two aspects affect the design process in ways that traditional methods struggle to address. Designers face not only technical challenges but also issues relating to how AI behaviour and outputs are conceptualised by the users. For example, users may struggle to understand why an AI system made a particular decision, leading to mismatches between their mental model of the system and its actual functioning. When designing systems that produce a virtually infinite number outputs through fuzzy, open-ended interactions, prototyping, sketching, and anticipating evolving behaviours become simply infeasible.

As a result, the literature is calling for a deeper understanding of the human experience with algorithms and the psychology of interacting with complex systems. Sundar, 2020 advocates for research designed to explore the effects of AI on human behaviour as AI systems become more autonomous and the boundary between human and machine agency becomes increasingly blurred. They argue that these should be designed with a clear understanding of how users perceive and respond to AI-driven decisions, urging a shift in research toward the symbolic and enabling effects of AI on human behaviour. Similarly, Jiang et al., 2023 discuss the need to address the tensions between automation and human agency, system uncertainty, and user confidence in AI applications. AI design should focus on improving transparency and user comprehension, enabling users to better understand AI decisions and interact more effectively with these tools. Amershi, Weld, et al., 2019 propose 18 design guidelines for AI-infused applications, which emphasise transparency in AI capabilities and performance, and the use of contextually relevant information and explanations to effectively manage user expectations.

Two sub-fields of ML appear as promising avenues for studying human-AI in-

teraction. The first is IML, which supports iterative experimentation by enabling designers and users to “play with” AI, gaining an intuitive sense of its capabilities. IML tools facilitate real-time engagement by allowing users to interact with and adjust models dynamically. This interactive process creates manageable test cases that help better understand complex interactions with AI (Q. Yang, Steinfeld, et al., 2020). An promising example is the study by Q. Yang, Suh, et al., 2018, where non-experts iteratively trained models using IML tools, revealing user trust levels, usability challenges, and common misconceptions, such as over-reliance on accuracy metrics. However, there seems to be a general lack of research using IML purely to investigate human-AI interaction, likely because building effective systems can be resource-intensive, and existing tools are still gaining traction in the field.

The second promising sub-field is XAI, which aims to improve the understanding of how ML models make decisions by providing interpretable outputs and explanation for model behaviour. A prime example is recent work by S. S. Y. Kim et al., 2023, which revealed how participants would use interpretability methods for purposes beyond just understanding AI outputs, such as calibrating their trust in the system, improving their skills, providing better inputs and giving developers feedback. Previous research leveraging XAI to study human-AI interaction also includes work from Ehsan et al., 2022, which explores the importance of embedding human-centred perspectives in explanations to cater to diverse human stakeholders who interact with AI at various stages. Q. V. Liao, Gruen, and Miller, 2020 also contributes to this perspective, proposing a question-driven framework that aligns XAI methods with user-centred needs. Their work highlights that user interactions with AI explanations extend beyond mere understanding, encompassing goals like assessing AI reliability, contextualising decisions, and providing feedback to improve system performance. Overall, it is clear from the literature that XAI is an essential component of human-AI interaction studies, focusing not only on model transparency but also on how explanations facilitate trust, effective use, and collaborative refinement of AI systems.

2.2 Interactive Machine Learning

IML 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-labelling misclassifications, providing and indicating representative samples and features, and setting preferences and thresholds (Amershi, Cakmak, et al., 2014; Dudley and Kristensson, 2018). 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 (Fails and Olsen, 2003; Arendt et al., 2019).

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 (Amershi, Cakmak, et al., 2014). Additionally, IML 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 (Arendt et al., 2019).

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 (Wallace et al., 2012; Holzinger, 2016; Berg et al., 2019). This approach also enables personalisation and fine-tuning, which is particularly relevant in creative applications. For example, performers could start with a system created by a designer and then customise it by incorporating their unique artistic style and preferences (Caramiaux and Tanaka, 2013). Alternatively, artists can train a model with their own data set and explore the output of the model until they believe the system has reached the desired level of performance based on their own expectations and needs (Sanchez et al., 2021).

An additional motivation for leveraging IML is the concept of “machine teaching”. With this term, the literature refers to the potential ability for this approach to encourage the exploration of strategies by users when training ML models (Simard

et al., 2017). In fact, by taking the role of a teacher teaching a machine how to perform a task, users are necessarily encouraged to build effective training strategies to iteratively refine model behaviour (Shneiderman, 2020). This process has been found to enhance user engagement and improve non-expert understanding of the data and the model through natural exploration (Sanchez et al., 2021).

Given the growing adoption and diverse applications of IML, relevant research has stressed the need for an increased understanding of end-user interaction design of these systems, considering interface design critical to the success of the iterative process (Amershi, Cakmak, et al., 2014; Corbett and Saul, 2018; Dudley and Kristensson, 2018). The difficulty in evaluating these tools and their interfaces is well-established, as the tight coupling between user and system makes the resulting mechanisms of co-operation and co-adaptation challenging to identify and interpret (Boukhelifa, Bezerianos, and Lutton, 2018). Still, considerable effort has been made towards better-informed system design following guidelines and heuristics to improve both user experience and model accuracy.

2.2.1 System Design

Addressing a lack of consolidated guidelines for IML system design is a review from Dudley and Kristensson, 2018, who proposed several solution principles following the four elements defined as: sample review, feedback assignment, model inspection, and task overview.

Not only is labelling data tedious and sometimes not considered worthwhile by the user, but it requires investing significant effort before noticeable change in the model (Wong et al., 2011; Groce et al., 2014; Ribeiro, Singh, and Guestrin, 2016). Notably, there appears to be an opportunity in the evaluation of interaction methods 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 labelling instances, such as feature selection and weight adjustment (Stumpf, Rajaram, Li, Burnett, et al., 2007; Porter, Theiler, and Hush, 2013; Dudley and Kristensson, 2018). However, these methods pose significant interface design challenges to avoid overwhelming the user with too many, or too advanced, machine-centric metrics, whereas data labelling remains the most popular method for end-user input (Hartmann et al., 2007; Amershi, Cakmak, et al., 2014).

Many possible causes of errors in ML fall under the categories of mislabelled data, feature deficiencies and insufficient data (Amershi, M. Chickering, et al., 2015). Several inspection methods allow the user to detect failures and their sources differently, including presenting all of the unlabelled data points sorted by their predicted scores for some class, and showing only the best and worst matches (Fogarty et al., 2008; Amershi, Cakmak, et al., 2014). A more effective presentation method evaluated by Amershi, Fogarty, et al., 2009 consists of summarising model quality while presenting low-certainty samples.

While most user studies on IML systems are time-limited and do not include termination conditions, visibility of global objectives and task status can address the inevitable point of diminishing returns reached in applications such as text classification systems (Groce et al., 2014). Charte et al., 2015 reveal that describing relatively simple strategies in the initial instructions and framing can greatly improve user consistency and understanding. In addition to an improved mental model, increased efficiency, and shorter sessions, strategies could also provide guidance to non-expert users when making task-level assessments even in experimental settings.

2.2.2 Model Inspection

The illustration of the current learned concept is considered a fundamental issue in end-user IML. Result visualisation methods can enable users to assess the quality of the model and inform how to proceed in training (Amershi, 2011). Different approaches have been evaluated for visualising predicted samples and labels. Users often change their strategies in candidate selection during the labelling process due to the observed behaviour in the model, and the likelihood increases significantly

when additional information about the data or classifications is available (Bernard, Hutter, et al., 2018; Dudley and Kristensson, 2018). However, if showing samples with their predicted labels enables users to easily assess the current state of the model, data-intrinsic properties like patterns or outliers emerge especially when labels are excluded (Bernard, Zeppelzauer, et al., 2018).

Evaluations of visualising samples from the training set alongside, or in separation from, predicted instances seem lacking. Direct visual comparison between trained and predicted samples might support the assessment of the current model state while attenuating biases introduced by exclusively presenting predicted labels. Equally overlooked is the interaction with trained samples, invariably considered ground truth and never to be re-labelled despite the changes in strategy of the user during the iterative process.

The visual representation of instances is highly dependent on data type and usually task specific, but some model- and data-agnostic methods can be effective (Bernard, Hutter, et al., 2018; Dudley and Kristensson, 2018). Pursuing the creation of novel interaction methods with ML models, Kapoor et al., 2010 developed and evaluated MiniMatrix, a system where users iteratively refine decision boundaries in a confusion matrix after each re-classification. The commonality and simplicity of confusion matrices as a visualisation method is leveraged mainly to identify and tune numerical parameter settings. However, results of the user study indicate that non-expert participants generally found the matrix valuable in providing insights about the structure of the classification problem, suggesting equally promising results in result visualisation.

A study by Q. Yang, Suh, et al., 2018 revealed that users who are not formally trained in ML tend to be more satisfied and trusting toward the learning results than their professional counterparts. This is because many non-experts build models exclusively for new insights on their data and disregard model accuracy or related metrics that are too complex and overwhelming (Beauxis-Aussalet and Hardman, 2014). Drawing inspiration from MiniMatrix, it is possible to speculate that confusion matrices are potentially more accessible to non-experts compared to these

measures.

2.2.3 Text Applications

As the vast number of digital documents continues to increase, automated text categorisation, information extraction, and summarisation have witnessed particular interest in the context of ML. Compared to different applications, there are numerous pre-processing tasks in the NLP pipeline that benefit from manual intervention, such as stop words selection, feature refinement, and dimensionality reduction (Bahrudin et al., 2010). Moreover, ground truth is arguably less defined in use cases such as qualitative research, where the goal is not to be accurate or objective because there is no single truth to inform data analysis correctly (Willig and W. S. Rogers, 2017). For these reasons, text applications seem particularly well-suited for the implementation of the IML process, and several systems have been developed and evaluated.

Abstrackr is a stand-alone annotation tool independent of its ML components aiming to semi-automate the laborious task of citation screening for systematic reviews in clinical research settings (Wallace et al., 2012). 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 colours within the text. Highlighting words or n-grams appears to maximise user perception of the features being exploited by the model and improve the understanding of the underlying function, including its deficiencies (Dudley and Kristensson, 2018).

ML is especially useful when data is large and complex, and the visualisations and interactions provided in IML applications should account for volume and dimensionality. A popular text visualisation technique, word clouds provide a high-level summary of text in a 2-dimensional space with font size proportional to the frequency of each word, but fail to provide context or structure necessary to inspect observed patterns (Chuang et al., 2012). In comparison, more advanced visualisations like Word Tree and DocuBurst employ interactive layouts to reflect semantic content and enable rapid querying and exploration of bodies of text (Wattenberg

and Viegas, 2008; Collins, Carpendale, and Penn, 2009). The cognitive advantages of spatial representations of information are well documented and can effectively support IML 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 (Andrews, Endert, and North, 2010; Endert, Fiaux, and North, 2012; Yu et al., 2015; El-Assady et al., 2017).

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 (Wall et al., 2018). The tool updates a control panel after each iteration of a Ranking Support Vector Machine model based on user preference, displaying information about the resulting changes in the attribute weights. Similarly, the prototype BrainCel features a spreadsheet where the user can select which points to edit, add to the training set, or predict (Sarkar, Jamnik, et al., 2015). 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 visualisation technique in text applications, the documented success of simple interfaces in enabling non-expert users to build ML models suggests a promising avenue (Sarkar, Blackwell, et al., 2014).

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 (Crowston, X. Liu, and Allen, 2010; Crowston, Allen, and Heckman, 2012; Tierney, 2012; Grimmer and Stewart, 2013; Lewis, Zamith, and Hermida, 2013; Liew et al., 2014; Muller et al., 2016). 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., 2018, such as a lack of understanding between disciplines, points to IML techniques as possible solutions.

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 (Gebreegziabher et al., 2023). The authors stress that, while in most domains, IML systems focus primarily on the optimisation 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 (Braun and Clarke, 2019), and entails the iterative exploration and review of new or existing patterns (Braun and Clarke, 2006).

2.3 Explainable Artificial Intelligence

Interpretability in ML has become a crucial area of research, commonly referred to as “Explainable AI”, especially as ML models are increasingly implemented in high-stakes environments such as healthcare, finance, criminal justice, and the military. Interpretability refers to the degree to which a human can understand the cause of a decision or predict the outcome of a model (Doshi-Velez and B. Kim, 2017).

The concept of interpretability is often discussed in contrast to the “opaque-box” nature of many modern models, where the internal workings are opaque sometimes even to the experts who designed them. Historically, simpler models like Decision Trees and Linear Regression were considered inherently interpretable because their decision-making processes could be easily traced and understood. However, as models have shifted towards more complex and non-linear architectures, achieving the same level of interpretability has become more challenging. For example, ensemble methods such as Gradient Boosting leverage the collective strength of multiple weak learners by combining their outputs to generate predictions. Deep Neural Networks consist of layers of non-linear transformations and mathematical functions that often involve millions or even billions of learned parameters to capture intricate patterns from large amounts of data. Transformers add another layer of sophistication by using self-attention mechanisms to process entire sequences of data simultaneously and learn contextual relationships more effectively.

A wide range of methods has been developed to address the challenge of in-

terpretability for these complex models. A key distinction made in the literature is between *global* interpretability, which seeks to provide an understanding of the general behaviour of the model, and *local* interpretability, which focuses on explaining individual predictions (Guidotti et al., 2018; Adadi and Berrada, 2018).

Global interpretability involves understanding which features influence model decisions, which is particularly helpful when ML models are used to inform population-level decisions, such as drug consumption trends or climate change (C. Yang, Rangarajan, and Ranka, 2018). Examples of global methods include inducing decision trees that approximate the model outcome (Craven and Shavlik, 1995; C. Yang, Rangarajan, and Ranka, 2018), quantifying the predictive power of individual input features at a global level (Covert, Scott M Lundberg, and S.-I. Lee, 2020), and utilising linear probes to generate confidence scores via flattened intermediate representations (Dhurandhar et al., 2018; Cesarini et al., 2024).

Local interpretability aims to explain why a specific decision was made for a particular input, which is useful when justifying why the model made a specific decision for a single instance (Adadi and Berrada, 2018). LIME, short for local interpretable model-agnostic explanations, creates a local surrogate model that closely mimics the prediction of the underlying opaque-box model for a single instance. Unlike the opaque-box model, these local surrogate models are transparent and interpretable. SHAP, short for Shapley additive explanations, explains the prediction of an instance by calculating the contribution of each feature to that prediction using concepts derived from coalitional game theory (Aechtner et al., 2022).

Methods can also be broadly categorised as *model-specific* and *model-agnostic* (Linardatos, Papastefanopoulos, and Kotsiantis, 2021). Model-specific approaches are tailored to particular types of models, drawing on their internal structures to generate potentially more accurate explanations. For example, in Neural Networks, the gradient associated with each input feature with respect to the output can be used to associate a score, or weight, to individual features for a given prediction. In contrast, model-agnostic approaches can be applied to any model regardless of its architecture. These methods are usually post hoc and generate explanations solely

based on the inputs and outputs of the model. LIME and SHAP are two examples of (local) model-agnostic methods that have gained widespread attention for their versatility and effectiveness, and represent the two most widely used XAI methods today based on the current literature in different domains (Aechtner et al., 2022; Cesarini et al., 2024; Salih et al., 2024).

2.3.1 Interpretability in Text Classification

Almost no inherently interpretable rule-based system is suitable for text applications due to the sheer size of the feature space (Ribeiro, Singh, and Guestrin, 2018). This fact has driven the development and evaluation of numerous post hoc interpretability methods for opaque-box models with trade-offs between interpretability, user trust, faithfulness, and computational efficiency.

A comprehensive evaluation of various post hoc interpretability methods for text using performance metrics by Atanasova et al., 2020 introduced a diagnostic framework to evaluate interpretability methods based on faithfulness, agreement with human rationales, confidence indication, and more. The study found that model-specific gradient-based methods generally outperform other model-agnostic methods for several Neural Network architectures, but also require more computational resources.

Another systematic evaluation of XAI methods in text classification using performance metrics and human evaluation (Cesarini et al., 2024) has recently provided several contributions to the literature. Methods focusing on individual feature importance were found to have higher fidelity compared to those that rely on general decision rules. Furthermore, while global explanations are perceived as more satisfying and trustworthy, they are also less practical than local explanations in many contexts.

Regarding local explanations, examples of classification instances have been shown to be a feasible vehicle to explain algorithmic behaviour (Caruana et al., 1999; Mikolov et al., 2013; Cai, Jongejan, and Holbrook, 2019), as this type of explanation has precedent in how humans sometimes justify actions by analogy (Lipton, 2018). However, the limitations of “explanations by example” have also

been widely discussed. For example, B. Kim, Khanna, and Koyejo, 2016 claim that, in order to construct better mental models and understand complex data distributions, explanations should also include what is *not* captured in the examples. Ribeiro, Singh, and Guestrin, 2018 argue that this is because these methods provide a trade-off: although each explanation is relatively easy to understand, they typically only capture the behaviour of the model within a specific, local region of the input space, which can lead to misunderstandings and poor approximations of the broader behaviour.

Traditional global rule-based interpretability techniques have been found to exhibit lower fidelity than local explanations (Cesarini et al., 2024) and, more importantly, do not increase task performance (Waa et al., 2021). Recent advances in LLMs have enabled efficient processing of large amounts of data that can capture complex patterns and nuances in text (Wei et al., 2022). Given these capabilities, there is an opportunity to enhance global interpretability techniques for text classification by leveraging LLMs to generate summaries of word importance weights from techniques like LIME. This approach could provide a more holistic view of model behaviour by aggregating feature importance across all instances in the training dataset, offering a coherent and comprehensive perspective on model decision patterns, instead of relying on limited or sample-based explanations.

2.3.2 User Studies

Doshi-Velez and B. Kim, 2017 established a baseline of evaluation approaches to interpretability techniques, proposing three major types: 1) application-grounded, where the explanations are implemented in an application and tested by a domain expert, 2) human-grounded, where experiments are run with non-experts within a simplified application, and 3) functionally-grounded, that does not involve users but rather performance metrics. In a recent systematic review on evaluating XAI, Nauta et al., 2023 reported that, among 312 papers published in the past 7 years at major AI and ML conferences that introduce an XAI method, 33% evaluated with anecdotal evidence, 58% applied quantitative evaluation, and only 22% evaluated with human subjects in a user study. Although numerous studies on interpretability tech-

niques often combine a variation of either application- or human-based evaluations with functionally-grounded evaluations (for example, in the very paper introducing LIME (Ribeiro, Singh, and Guestrin, 2016)), this work has specific interest in how previous research has designed user studies to draw meaningful conclusions on interpretability techniques.

A popular approach to evaluate interpretability in user studies involves a direct review of model explanations with end users for their subjective feedback. In fact, many papers report measurements of user understanding of explanations (Lage et al., 2019; Poursabzi-Sangdeh, Daniel G. Goldstein, et al., 2021b) or self-reported trust (Nourani et al., 2019; Papenmeier, Englebienne, and Seifert, 2019) as a proxy for usefulness and interpretability. However, recent work has recently revealed significant limitations of these methods for evaluating model explanations involving subjective feedback. For example, it has been shown to be possible to fool users in accepting wrong decisions (Schneider, Meske, and Vlachos, 2021) and manipulate user trust (Lakkaraju and Bastani, 2020) with misleading explanations that are not faithful to the opaque-box model. Furthermore, human judgment ratings can include user cognitive biases toward visual appearance or completeness of saliency maps that result in incorrect ratings (Mohseni, Block, and Ragan, 2020). In a study from Buçinca, Lin, et al., 2020, participants reported a higher preference and trust in an AI decision support system with images that used explanations, but the explanations did not translate to an improved performance in making accurate predictions.

To address the shortcomings of these methods, Mohseni, Block, and Ragan, 2020 proposed a “human-grounded benchmark” using human-attention data to compare generated saliency maps in images or text against user annotations of regions or phrases most representative of the target topic. However, despite offering a more objective and reliable evaluation method, a reported limitation of the benchmark is the significant cost of manual data annotation. Instead, an alternative evaluation method is based on a previously proposed metric to assess the interpretability of a system according to which, if users truly understand how the system functions, they should be able to accurately predict its output (Muramatsu and Pratt, 2001).

Waa et al., 2021 reported on a user study asking participants to predict the outcome of a personalised advice system for the self-management of diabetes to compare rules-based to example-based explanations. B. Kim, Khanna, and Koyejo, 2016 designed a similar predictive task in which participants were asked to predict the outcome of an image classifier based on “prototypes” (representative examples of typical instances of model behaviour) and “criticisms” (examples that differ significantly from prototypes). Buçinca, Lin, et al., 2020 included a proxy task where participants focused on predicting AI recommendations from an image decision support system based on example-based and general explanations. Alqaraawi et al., 2020 developed a user study on image classification in which participants were shown a collection of true positive, false negative and false positive images from a trained CNN with or without LRP saliency maps, and then asked to predict the outcome of the same model on an additional image close to the examples in terms of vector distance. The evaluation method used in these studies is usually defined as “forward simulation” (Belle and Papantonis, 2021), and has not yet been applied to interpretability in text classification.

2.4 Summary and Discussion

As AI becomes increasingly embedded in everyday applications, human-AI interaction is becoming a focal point in HCI research. Recent literature calls for a deeper understanding of how users experience and interact with AI systems, with researchers emphasising transparency, comprehension, and alignment with human cognitive models to foster better user interaction with complex and unpredictable systems.

Within this field, IML and XAI are two areas where HCI research can take place to improve human-AI collaboration and understanding. The first allows users to interact with and adjust ML models iteratively by enabling dynamic experimentation from frequent, iterative interactions with core features of ML. XAI seeks to make the decision processes of ML models interpretable, and in doing so, also touches on additional aspects, such as performance assessment, trust calibration,

and mental model adjustments.

Section 2.2 shows that the existing literature on IML is largely focused on design guidelines and implications. In contrast, there is a general lack of research specifically utilising IML as a means to investigate human-AI interaction dynamics. For example, how do users understand and interpret the purpose of ML within the context of their interactions? How do mental models, perceptions, and decision-making strategies evolve over time as users engage in the iterative cycle of IML? How does the integration of IML affect the depth and scope of engagement with the data? Answering these questions would help achieve the deeper understanding of the psychology of interacting with complex systems called by studies such as those of Sundar, 2020 and Jiang et al., 2023 and ultimately advance the design of AI systems that align better with human cognition and behaviour. This observation motivated the formulation of **RQ1-3** described in Section 1.1.

Chapter 3 addresses a research gap in the missed opportunity to utilise IML to uncover actionable insights in human-AI interaction. Although the findings provide several design implications on the model inspection and feedback assignment phases of the IML iterative cycle, its main contribution is an improved understanding of non-expert perceptions and misconceptions of ML models. Chapter 4 further builds on this to examine interactions in more depth through sustained engagement, contributing to the literature with a first-hand account of evolving re-labelling and model inspection strategies, and the effects of integrating ML into the analysis of data.

The literature reviewed in Section 2.3 shows that previous work has evaluated interpretability techniques by measuring one or a combination of performance metrics (Atanasova et al., 2020; Cesarini et al., 2024), user understanding (Cesarini et al., 2024), self-reported trust (Nourani et al., 2019; Papenmeier, Englebienne, and Seifert, 2019; Cesarini et al., 2024), and human-grounded benchmarks (Atanasova et al., 2020; Mohseni, Block, and Ragan, 2020). Instead, Chapter 4 follows the reasoning of Muramatsu and Pratt, 2001 to measure outcome prediction to adopt the evaluation method used in related work on different task domains (B. Kim, Khanna,

and Koyejo, 2016; Alqaraawi et al., 2020; Belle and Papantonis, 2021; Waa et al., 2021) to extend the findings to text classification.

The contribution of Chapter 4 in answering **RQ4** thus addresses a specific methodological research gap where interpretability techniques in text classification have not yet been evaluated through model outcome prediction. This gap is particularly significant because some alternative evaluation methods have been shown to either fall short in incorporating the user-centred evaluations encouraged in the literature (Abdul et al., 2018; D. Wang et al., 2019; Q. V. Liao, Gruen, and Miller, 2020), or make use of misleading measures that fail to account for biases in human judgement (Lakkaraju and Bastani, 2020; Mohseni, Block, and Ragan, 2020; Schneider, Meske, and Vlachos, 2021). As argued by Nauta et al., 2023, interpretability is multi-faceted, and a single metric cannot capture the effectiveness of a technique. Therefore, the domain of text classification can benefit from the application of an already established evaluation method that can complement the existing methods used in the literature so far.

Chapter 3

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

This chapter introduces TACA, a fully-functional IML application designed and developed to assist users in QDA and deployed to run on Windows and MacOS entirely offline to preserve the confidentiality of the data used. It then presents an evaluation of human-AI interaction enabled by TACA in the context of thematic analysis. Thematic analysis is a QDA research method used to identify, analyse, 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.

As discussed in Section 1, QDA is particularly well suited for research in human-AI interaction because the lack of ground truth means that data is ambiguous and thus allows for a more nuanced exploration of how users navigate through different perspectives. Additionally, IML allows users to steer the model according to personal interpretations and emerging insights in real-time, an important aspect of

QDA. However, while QDA serves as a valuable application domain for this study, the main interest in this work lies in user interaction with ML. It remains unclear how non-expert end-users understand and interact with these systems, including potential biases involved in the iterative process, so this work aims to address this particular research gap.

Section 2.2 shows that the literature on IML in HCI is largely focused on design guidelines and implications. The chapter explores these aspects, specifically by evaluating a novel data aggregation technique based on word frequency to address reported limitations of IML systems, but the main goal is to answer the following research questions described in Section 1.1:

- **RQ1: How do non-expert users perceive ML when analysing ambiguous data?**
- **RQ2: How do non-expert users' perceptions of ML influence their interaction with it?**

The decision to evaluate specifically *non-expert* interaction was based on the premise that, besides reducing the requirement for extensive data sets, a significant advantage of IML is that model refinement can be driven by users who lack specialised knowledge in ML (Ware et al., 2001). Additionally, as the general population is more frequently exposed to AI systems, understanding how non-expert users perceive and interact with ML becomes especially valuable and can reveal insights that are increasingly relevant.

The answers to the two RQs were revealed through thematic analysis of the responses of 20 participants in a semi-structured interview (see Appendix A.5) conducted after the participants spent time interacting with TACA. The findings were also supported by a quantitative analysis of automatic interaction logs aimed to capture user behaviour.

The work in this chapter is to appear as a paper in the ACM Conference on Computer Supported Cooperative Work (CSCW) in 2025:

Federico Milana, Enrico Costanza, Mirco Musolesi, and Amid Ayobi.

“Understanding Interaction with ML through a Thematic Analysis Coding Assistant: A User Study”. *Proceedings of the ACM on Human-Computer Interaction (CSCW)*

The following section provides additional details on the context of QDA in this work.

3.1 Qualitative Data Analysis as an Application Area for Interactive Machine Learning

The literature on IML 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 (Amershi, Cakmak, et al., 2014). 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. In addition, evaluations of smart systems through a controlled yet ecologically valid study requires experimental tasks to be engaging and enjoyable to motivate participants and provide meaningful discussion points (Kittley-Davies et al., 2019).

Prior work highlighted the potential to apply ML to QDA (Chen et al., 2018; Gebreegziabher et al., 2023). 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 (Lazer et al., 2009). This issue may be addressed by applying IML techniques to help mitigate the impact of lower system accuracy. Numerous applications implementing the IML 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 (Wallace et al., 2012; Sarkar, Jamnik, et al., 2015; Yu et al., 2015; El-Assady et al., 2017; Wall et al., 2018; Q. Yang, Suh, et al., 2018; Gebreegziabher et al., 2023).

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 re-

quire a large quantity of labelled 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 (Lewis, Zamith, and Hermida, 2013), prompted to consider a different approach to code automation. Instead of automating the coding process, there is a clear opportunity to assist researchers in reflecting on their completed analysis by providing additional automated coding suggestions.

3.2 The Thematic Analysis Coding Assistant (TACA)

To enable a user study on applying IML to QDA, TACA was developed as a fully functioning GUI desktop application designed to assist specifically with the coding phase of thematic analysis. TACA was deployed as an executable to run on Windows (minimum version: 8) and MacOS (minimum version: 10.9). 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 the 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-labelling sentences from one theme to another), re-train the ML classifier to interactively refine it, and, in so doing, customise it to produce coding suggestions that align better with the user's analysis.

Because qualitative data can often be confidential and researchers may 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 Microsoft Word or popular QDA software NVivo¹, MAXQDA² or Dedoose³ as the original coding environment source. After importing the data, users can define a list

¹<https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home>

²<https://www.maxqda.com/>

³<https://www.dedoose.com/>

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 analysed.

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 grey are the user-coded sentences, while those predicted by the model appear in blue, seen in Figure 3.1. 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.

The screenshot shows a software interface titled 'Text' at the top left. Below the title are three tabs: 'Codes', 'Keywords ▾', and 'Confusion Tables ▾'. The main area displays a transcript of a conversation between two people, P1 and IV2. The transcript is divided into several sections by vertical lines. Some sections are highlighted in grey, indicating user-coded sentences. Other sections are highlighted in blue, indicating sentences classified by the model. To the right of the transcript, there are several vertical bars of different colors (grey, blue, and light blue) representing different themes. Each bar has a label next to it: 'practices', 'study vs product', 'practices', 'system perception, system use', 'system perception, system use', and 'social, system use'. The transcript content includes:

- 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.
- 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.
- P1: 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.
- IV1: Yes, makes sense. What do you think about how the system knows when to order?
- 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.
- IV2: Did the delivery times always work?
- 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.

Figure 3.1: Text page showing highlighted user-coded sentences in grey 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 3.2, 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 previous research (Dudley and Kristensson, 2018), it was hypothesised that giving salience to indica-

tive keywords accelerates the assessment process. Therefore, the terms are sorted by frequency, where each column is a theme, shown in Figure 3.3. 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 grey for training samples and blue for predicted. Unique to all Keywords Tables is the re-labelling 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. Keywords are used as handles for groups of sentences to enable the user to re-label multiple data points, or sentences, at once.

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)			x engagement (20)	autonomy (59)	like (13)
busy (15)			x expectation (16)	awareness (32)	not working (8)
change of practice (29)			xental model (74)	battery (15)	sustainability (51)
consumption (3)			xace of mind (1)	cancelling (19)	working (25)
cook (63)			xice (0)	check (16)	works (7)
cooking (53)			xovenance (6)	checking (9)	
diet (2)			xuality (31)	contextual (7)	
eat (85)			xlf-blame (17)	control (77)	
eating (72)	researcher in the loop (1)		xuperfluous (11)	delay (55)	
eating pattern (15)	social (48)		xtrust (36)	delivery (208)	
exceptions (17)			xunderstanding (15)	disconnection (2)	
existing practice (28)			xuseful (0)	fixing (12)	
fattening (3)				lazy (20)	

Figure 3.2: 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 visualisations in

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)			X every (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)			o (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)	

Figure 3.3: All Keywords Table page showing the most frequently occurring terms for each theme.

progressive analysis by Stolper, Perer, and Gotz, 2014. 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-labelled 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 3.4, 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.

3.2.1 Implementation Details

TACA was implemented mostly in Python to leverage the availability of ML libraries. The PyQt⁴ 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 NLP library NLTK⁵, and stop words defined in the same library are ex-

⁴<https://github.com/pyqt>

⁵<https://www.nltk.org/>

"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)
..... (4) (3) (2) (2)

Figure 3.4: Confusion Table page showing the most frequently occurring terms for each confusion matrix quadrant of the selected theme.

cluded. 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 set⁶. 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 XGBoost⁷ 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, ClassifierChain from scikit-learn⁸ was used to 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, the MLSMOTE data augmentation algorithm for multi-label classification (Charte et al., 2015) is applied before training the chains of classifiers on the labelled embeddings. Following training, all uncoded sentences in the text are predicted using a confidence threshold of 95%, based on the average

⁶<https://nlp.stanford.edu/projects/glove>

⁷<https://xgboost.readthedocs.io/>

⁸<https://scikit-learn.org/stable/modules/generated/sklearn.multioutput.ClassifierChain.html>

of binary predictions from the classifier chains.

When the user imports the text containing coded sentences, the data set is split 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 score⁹ 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 (Shen et al., 2020). Additionally, TACA does not handle ambiguous data explicitly the way previous research in noisy data in ML proposed (Schlimmer and Granger, 1986; Raychev et al., 2016; S. Gupta and A. Gupta, 2019), 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 labelling process, as well as reviewing the coding suggestions generated by the classifier.

Multithreading enables TACA to load every page independently and simultaneously while prioritising 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-labelling 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. The code is released as open source¹⁰.

⁹The F1 score is a metric that combines precision (how many of the positive predictions are correct) and recall (how many actual positives the model identifies) into a single number ranging from 0 to 1. A higher F1 score means that the model is better at balancing these two aspects, leading to more accurate and reliable predictions overall.

¹⁰<https://github.com/fmilana/tacodingassistant>

3.3 Study

The study was reviewed and approved by the Ethics Committee of UCLIC (Project ID No: UCLIC_2022_004_costanza). All participants were volunteers and provided informed consent before taking part in the study (see Appendix A.3).

3.3.1 Participants

20 participants were recruited from Prolific¹¹ (an online crowd-sourcing recruitment platform), a psychology and language science participant pool at UCL, 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 3.1. The gender imbalance among participants reflects the demographic composition of the fields from which the recruitment was conducted, particularly in areas where qualitative research is commonly undertaken. Previous research in psychology and social sciences has found that gender can sometimes influence perspectives, cognitive styles, and analytical approaches (Sternberg and L.-f. Zhang, 2014; Alalouch, 2021). Although the primary objective of this work was to expose a representative sample of QDA practitioners to IML, a promising future direction is to conduct more gender-balanced and gender-focused user studies in this area.

3.3.2 Procedure

Participants were given TACA 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 A.4). 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 analysed for publication, ranging from a few months

¹¹<https://www.prolific.co/>

Table 3.1: User study participants information.

ID	Age	Sex	Occupation	Area of research	QDA		Recruited from
					experience	Data used (years)	
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

to a few years prior to the study. To facilitate recruitment, a collection of 21 reviews of restaurants published in the newspaper *The Guardian*¹² between 2022 and 2023 was distributed to manually code by participants who did not have their own data sets available for the study. Restaurant reviews were chosen because the topic did not require specialised knowledge, and the reviews were expected 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 analyse 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.5). 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 recordings of their screen could be revisited later to contextualise 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.

¹²<https://www.theguardian.com/food/restaurants+tone/reviews>

3.3.3 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 IML and TACA. Audio recordings from the interviews were transcribed verbatim and then analysed using inductive thematic analysis (Braun and Clarke, 2019; Byrne, 2022). An early phase was focused on the explicit meaning of the participants' accounts through the familiarisation 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, where the initial themes were revisited and modified 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 (see Appendix B.1), discussed in the following section.

3.4 Findings

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

3.4.1 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-labelled 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-labelling 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-labelled at least one data point re-labelled, 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 of sentences into themes), 2.0 ($SD = 3.5$) in the Predict Keywords Table (i.e., they corrected the classifications of the ML model 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-labelled 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-labelled 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-labelled data points and participant demographics, i.e. age, sex, occupation, field of study/research and

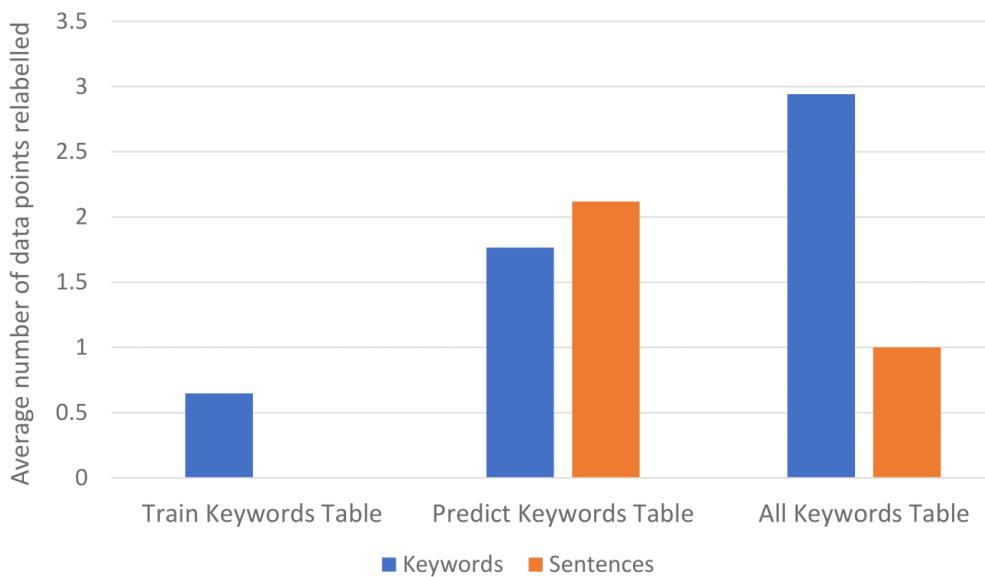


Figure 3.5: Average number of data points re-labelled in each table.

QDA experience.

3.4.2 Evaluation Strategies for Model Inspection and Reflections on Machine Learning Output

Following the initial quantitative analysis of interaction data, this section explores the evaluation strategies participants employed, reflecting on the output of the model 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 analyse 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 summarising results by “*synthesising*” (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 analyse 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. Recognising 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 predictions of the model 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 analysing 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 organise “*better themes and sub-themes*” (P10) overall.

3.4.3 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-labelling. 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 organise 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-labelling 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 analysing keywords, since that’s a big part. It’s got a whole section with keywords*” (P15).

3.4.4 Perception of the Machine Learning Model

Having identified the specific misconceptions around the use of keywords and sentences in the batch re-labelling process, this section now shifts attention 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 categorising the codes*” (P9). Still, many participants viewed the model as an external source offering objective advice, a perspective reflected in numerous observations, such as: “*It’s like an external source that’s analysing 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 an indication of 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 coding of the model: "*The model must have some reasoning for categorising these words into the false negatives.*" (P9)

Nevertheless, there was still value found in the Confusion Tables when evaluating the model. One approach involved "*forming an opinion depending on the quality of the false positives*" (P10): "*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). Analysing "*what [the model] is suggesting and maybe what it's also not suggesting*" was an effective strategy "*to see what the model thinks*" (P5).

3.4.5 Personal Blame for Poor Model Performance

The perception of an external source of objective advice naturally revealed a second theme 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 behaviour of the model 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.

3.4.6 Perceived and Anticipated Use of Interactive Machine Learning in Qualitative Data Analysis

Finally, this section considers the wider context of integrating IML in existing data analysis workflows. Participants recognised that analysing 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 organising 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 behaviour*” (P9).

Nevertheless, there was a clearly perceived distinction between the researcher and the tool. “*I think the role of the tool is to organise 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 recognised that, instead of replacing the researcher, TACA would complement them by cross-checking data, evaluating saturation, organising 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 realised 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).

3.5 Discussion

Through an analysis of interaction logs and semi-structured interviews, the situated account provided described how participants analysed qualitative data using an IML system. The findings demonstrate TACA as a functioning and usable tool to identify the benefits and challenges of enabling non-ML experts to engage with IML. These findings have implications that extend beyond the scope of the tool and can be applied to various domains outside QDA. The following three sections of the discussion focus on how IML 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 explored to facilitate the IML cycle.

3.5.1 Supporting Reflexivity with Interactive Machine Learning

Participants recognised and valued the process of reviewing their own analysis, identifying patterns, gaining deeper insights, and re-interpreting findings using TACA. The advantages of using IML in QDA reported by the participants confirm the results of Gebreegziabher et al., 2023, which highlight the importance of the ability of researchers to refine and evolve their coding frameworks in collaboration with AI tools. Marathe and Toyama, 2018 found that researchers desire automation only after having developed a codebook and coded a subset of data, particularly

by extending their coding to unseen data, and most of the participants in the study confirmed this in the interviews. However, the benefits of IML 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 IML states that result visualisation techniques can enable users to assess the quality of the model and inform how to proceed in training (Amershi, 2011). Because the study involved subjective, ambiguous data with no objective ground truth, participants utilised result visualisation 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, 2017 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”. During the 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 (Fontana, 2004; Jootun, McGhee, and Marland, 2009; Braun and Clarke, 2019; Holmes, 2020). Reflexivity allows researchers to critically assess their own influence on the research process and outcomes, and in the 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 miscategorised, recognising their own perspec-

tive 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 behaviour suggests that IML 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-centred 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, IML systems should be designed to promote critical engagement with data by providing clear and insightful feedback on both generated classifications and manually labelled data samples to allow for comparisons in a similar fashion to Confusion Tables in TACA.

3.5.2 Balancing Objectivity and Subjectivity in Interactive Machine Learning

The study results emphasise reflexivity as a key benefit of IML, 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, the participants often considered these belonging to an equally valid, if not better, interpretation of the data. This perception could also explain why participants re-labelled fewer data points than expected and re-trained the model only a limited number of times.

A recent study by Q. Yang, Suh, et al., 2018 revealed that non-experts are generally more satisfied and trusting toward the outcome of ML compared to their professional counterparts, which can explain why participants almost always blamed themselves when recognising 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. Supposedly, this could also have been influenced by the underlying goal of improving the accuracy of the model through the process of re-labelling 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.). This terminology was adopted 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 (Beauxis-Aussalet and Hardman, 2014).

In fact, while recent work has found that non-experts often struggle with the standard terminologies and structural design of confusion matrices (Shen et al., 2020), most participants in the study clearly understood how to interpret the confusion tables and rarely required guidance in the interviews. Still, the terminology used might have inadvertently contributed to the perceived objectivity of the model, despite the fact that most participants recognised that the model was trained on their own, subjectively labelled 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 (Alvarado and Waern, 2018). In the study, participants clearly recognised that re-labelling 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 con-

structed (Miceli, Schuessler, and T. Yang, 2020). The reason might be that, in the ML academic communities, contributions are determined by the modelling 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 (Miceli, Schuessler, and T. Yang, 2020). Subjectivity in ML is also manifested in the processes of meaning-making, modelling choices, and data idiosyncrasies (Javed et al., 2021; Waseem et al., 2021), so, while the participants’ perception that ML is intrinsically “objective” is not surprising, it should certainly be challenged.

Implementing interpretability techniques in TACA was beyond the scope of this work, as the aim of the study was to observe how participants would interact with the basic version of an IML tool without introducing additional complexity. However, the findings reveal a need for transparency in IML tools to help users understand the inherent limitations of ML models. Transparency has been found to encourage users to provide more labels (Rashid et al., 2006) and with higher accuracy (Rosenthal and Dey, 2010). 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 (Amer-shi, Cakmak, et al., 2014) and, more notably, calibrate trust in the model, especially for non-experts (Ribeiro, Singh, and Guestrin, 2016; Dudley and Kristensson, 2018; Ayobi et al., 2021). 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 (Severes et al., 2023), and the results support the need of these recommendations for the use of explanations not just for ML practitioners in model diagnostics.

3.5.3 Understanding Perceptions of Machine Learning through User Interface Features

Building on the insights into the balance between objectivity and subjectivity in IML, this section investigates in more depth how specific UI features influenced

participants' perceptions and interactions with the model. The frequency-based keywords in the Confusion Tables and Keywords Tables were intended to provide participants with meaningful insights about the current state of the model, and also allow them to efficiently manipulate the large and high-dimensional data set for re-training the model, which is normally challenging. The findings regarding keywords are specific to the UI of TACA and not universally applicable to all uses of IML. 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 the participants' experience with TACA.

The study evaluated the presence of confusion matrices displayed as tables containing representative samples in the model inspection phase of IML. 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. The findings revealed that most participants naturally took different approaches to explore the output in the Keywords Tables. Most of the 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 modify the behaviour of algorithms to suit their needs and preferences (Alvarado and Waern, 2018). The 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-labelling individual sentences. Most of the keyword were re-labelled 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-labelling multiple, semantically similar samples can be effective in the feedback assignment phase of the IML cycle, reducing the significant effort required to label data points in large data sets (Wong et al., 2011; Groce et al., 2014; Ribeiro, Singh, and Guestrin, 2016).

However, participants were generally hesitant to interact with the Keywords Tables and re-labelled fewer data points than expected. Dragging and dropping whole keywords was considered by some too big of a move. This behaviour reveals a limitation of keywords as implemented in TACA, as the interface does not allow users to undo changes and lacks mechanisms for previewing potential modifications before committing to them. In addition to enabling users to revert to a previous stage of the model, it can be hypothesised that anticipating the resulting changes before re-training could have increased the participants' confidence in re-labelling data points. Interfaces for feedback assignment in IML requires the most careful design in terms of both elements and interaction techniques (Dudley and Kristensson, 2018), and visualising anticipated changes can introduce transparency in the system, greatly affecting the quality of the response elicited from users (Amershi, Cakmak, et al., 2014).

IML applications often assume that users possess domain knowledge, which is crucial for accurate model inspection and feedback (Amershi, Cakmak, et al., 2014). Therefore, the initial intention was to recruit only participants who had their own coded data to analyse, but due to recruitment challenges, it was decided to provide newspaper restaurant reviews to those who did not have any data available. Eventually, only 5 out of 20 participants analysed 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. This would explain why this group moved fewer keywords on average. Additionally, the group that used their own data seemed to be slightly more critical of the model suggestions, preferring to re-label more sentences that were classified by the model rather than those that were manually labelled in their own analysis.

The decision to use keywords as a data aggregation technique also introduced misconceptions about keywords and sentences. Some participants believed they were re-labelling individual words instead of the sentences that contained them and that, consequently, TACA operated at the word level. This can be recognised as an additional limitation of using keywords as handles, but the need for users to manipulate data sets that are challenging to represent and summarise due to their size and dimensionality is an acknowledged existing challenge of IML. It is possible to speculate that the limitation observed in this 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-labelling.

Other applications of ML on text, such as sentiment analysis and information retrieval, could benefit from aggregation to support the IML cycle. The findings suggest that this approach can be effective, but it is essential to design around these features carefully to avoid misinterpretation. IML 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, visualisations 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-labelling process.

3.6 Conclusion

This chapter reported on a user study where 20 participants without prior ML experience used TACA, a novel IML application designed and developed to enable

the study. The focus was on thematic analysis as a practical application of IML: thematic analysis involves individual interpretation of ambiguous data and hence it is suited for and can benefit from the iterative customisation of the model. The 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 provided to them (newspaper restaurant reviews), if they did not have data available.

TACA was effective in exposing the participants to ML and apply it on their data. Participants recognised the value of incorporating ML in the thematic analysis workflow as the presence of coding suggestions encouraged a more critical analysis of data. Keywords and Confusion Tables, features of the TACA UI, also supported the model inspection and feedback assignment phases of the IML cycle but introduced misconceptions around the mental model of the tool. Finally, the 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.

IML has significant advantages over conventional ML, but the success of this alternative approach is strongly dependent on our understanding of user perception and interaction with ML models. Hopefully, this 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.

The limitations of this study include the unfamiliarity with the data of the participants using the restaurant reviews, the infrequent interaction with the ML model, and the lack of genuine motivation to extend their analysis. The next chapter presents an autoethnography on TACA to address these limitations and further investigate the application of IML to QDA.

Chapter 4

An Autoethnography on the Thematic Analysis Coding Assistant

Despite the growing interest in using ML-driven tools, such as TACA, in qualitative analysis, there is limited research exploring how ML is experienced by users first-hand over extended periods. The short-term nature of user studies at the intersection of HCI and AI means that behaviour can be measured and generalised but does not fully capture how it evolves as people continue to interact with ML. Furthermore, since user studies often take place in siloed settings, participants are almost never motivated by long-term investment in the outcome of their interactions.

Moreover, the key findings of the study reported in Chapter 3 included that participants perceived the model as objective and authoritative, attributing any issues with performance to their own input data. Consequently, the participants engaged only to a limited extent with the IML features of TACA. In contrast, this chapter reports on a follow-up study designed to place more emphasis on these IML features: an autoethnography conducted to answer the following question:

- **RQ3: How can IML be used to support the analysis of ambiguous data?**

The purpose of the autoethnography is threefold: 1) to collect personal reflections on how the user interface and feedback mechanisms of TACA influenced decision-making, re-labelling strategies, and overall workflow; 2) to explore in greater depth personal perceptions, thoughts, and emotions during the interaction with the ML model; and 3) to compare the results obtained from extending the

analysis of the interviews on TACA with the outcomes of the initial manual analysis reported in the previous chapter. In more general terms, the goal is to uncover nuances in the interaction that may not have been evident during the user study, as well as additional insights that leverage the roles of developer, researcher, and participant with experience in both HCI and AI.

This work appears to be the first autoethnography on IML. Addressing this research gap is important because the iterative cycle of IML is heavily influenced by critical thinking, decision making, and perception of the model, as revealed in the user study with TACA. A detailed exploration of how one's own biases, expectations and expertise can provide valuable insights for designing more effective interactions with ML that better account for user experience. Additionally, while prior work has applied autoethnography at early design stages, this work contributes by demonstrating how autoethnography can generate new insights even after a system has been designed and evaluated.

The work in this chapter has been presented as a paper under review for the ACM Conference on Intelligent User Interfaces (IUI) in 2025:

Federico Milana, Enrico Costanza, Mirco Musolesi, and Amid Ayobi.

“Understanding Interactive Machine Learning through an Autoethnography of the Thematic Analysis Coding Assistant (TACA)”. *Proceedings of the ACM on Human-Computer Interaction (IUI)*

The following section introduces self-study research methods, outlining how approaches like autoethnography provide a framework for understanding complex, lived experiences in HCI, and how these can be applied to AI systems.

4.1 Self-study Research Methods

4.1.1 Autoethnography

Recent epistemological shifts in social sciences suggest that facts and truths are intrinsically shaped by the paradigms and vocabularies used (Rorty, 1994; Kuhn, 1997). This philosophy rejects the idea of universal narratives in favour of more localised and contextual understanding (Lyotard, 1984; De Certeau and Rendall,

2004), emphasising the complex role stories play in shaping morals, ethics, and sense making (Bochner, 1984; Fisher, 1984; Bochner, 1994; Tony E Adams, 2008). In line with this perspective, autoethnography is a research method that combines ethnographic fieldwork with personal narrative to describe and systematically analyse personal experiences (Ellis, Tony E. Adams, and Bochner, 2011). Increasingly used within qualitative research, autoethnography differs from traditional ethnography by prioritising the researcher's own personal narrative, providing insider perspective on the subject matter.

According to Ellis, Tony E. Adams, and Bochner, 2011, the three fundamental pillars of autoethnography are 1) reliability, 2) validity and 3) generalisability. Reliability refers to the credibility of the narrator, whether the experiences described are plausible, genuinely believed, and not distorted into fiction (Bochner, 2002). Validity refers to the extent to which the work evokes a sense of lifelike, believable experiences, and emphasises the coherence of the “story” and how it relates to its readers (Ellis, 2004). In the absence of large random samples, generalisability is instead tested by the readers, who determine whether the story resonates with their own experience or illuminates unfamiliar cultural processes (Flick, 2020).

4.1.2 Self-study in Human-Computer Interaction

Self-study is not new in HCI (O’Kane, Y. Rogers, and Blandford, 2014; Desjardins, Tomico, et al., 2021). In general, however, while traditional autoethnographies focus on drawing lessons about social and cultural “texts, experiences, beliefs and practices” (Tony E. Adams, Ellis, and Jones, 2017), many HCI autoethnographies shift the emphasis to understanding how the features of a design shape and reflect interactions with technology (W. Gaver and F. Gaver, 2023). Researchers have applied different autoethnographic methods at various stages of the design process, such as informing user study design (D. Jain et al., 2018), testing preliminary prototypes (Bergman and Haitani, 2000; Buchenau and Suri, 2000), and as part of iterative design cycles (Neustaedter and Sengers, 2012; Pijnappel and F. ’. Mueller, 2014; S. Jain and Wallace, 2019). Adopting the role of participant has allowed researchers to gain a deeper understanding of user experiences that may be difficult or impossi-

ble to access using traditional methods (Neustaedter and Sengers, 2012; Desjardins and Ball, 2018). For example, researchers who become participants themselves are able to empathise with users, as seen in studies where they experience frustrations and limitations of a technology first-hand (Buchenau and Suri, 2000; Höök, 2010; O’Kane, Y. Rogers, and Blandford, 2014).

Self-study methods have particular value for artists and those engaging in artistic and arts-based research who want to offer deeply personal insights into design processes, exploration, and user experience (Musgrave, 2019; Bartleet, 2021; Ecclesia, 2023). For example, Mainsbridge, 2022 offered a personal account of working with motion-sensing interfaces the underlying qualities and meanings of performance actions in live and recorded contexts. Similarly, Väkevä, Mekler, and Lindqvist, 2024 detailed intimate insights into the trajectory and emotional qualities of personally meaningful and transformative videogame experiences. These studies show how self-study can illuminate the subjective and experiential dimensions of research, offering nuanced perspectives that might be overlooked in traditional approaches.

4.1.3 Self-built systems

The use of first-person research methods that involve self-built systems has also been explored in HCI. Some authors make a distinction between autoethnography and *autobiographical design*, in that the latter is a form of autoethnography that is located within the design cycle of a system (Neustaedter and Sengers, 2012; Desjardins, Tomico, et al., 2021; Bang et al., 2024). In reality, the boundary is not always distinct, as it is not uncommon for autoethnographies to conclude with a specific design guidelines section or a concrete set of opportunities for design (Aoki, 2007; Höök, 2010; Pijnappel and F. ’. Mueller, 2014; Lucero, 2018).

W. Gaver and F. Gaver, 2023 provide an “autoethnographical account” of their experiences using two self-built communication devices using coloured light and discuss the features that have proven important in mediating feelings of connection. They argue that the strength of this approach lies in their direct, long-term access to their lived experiences with the system, offering a first-person perspective instead of

relying on third-person observations or descriptive data. Similarly, autobiographical design leverages long-term usage to deeply understand the effects of the system on real practice as opposed to novelty effects (Neustaedter and Sengers, 2012). This level of understanding is difficult to obtain with other research methods, as it requires approaches that capture real everyday usage, which is often missed in short-term evaluations or siloed studies.

Seminal work discussing “tensions” in first-person research methods on self-built systems comes from Desjardins and Ball, 2018. Key challenges include defining genuine needs, balancing participation from others, managing intimacy and privacy, navigating the reflexive relationship between design and research, and addressing the complexities of reporting from the dual perspective of researcher and participant. To address these tensions, the authors suggest emphasising sincerity and transparency in documenting motivations, carefully managing collaboration and authority, and being inventive in reporting and reflecting on personal experiences.

4.1.4 Self-study in Artificial Intelligence

Only recently have some studies explored the application of first-person research methods to ML systems. Arvidsson and Noll, 2023 present an autoethnography of the authors’ attempt to build an ML system to mitigate discrimination in asylum law decisions. Through their experiences, they report how they realised that, instead of eliminating discrimination, the system simply shifted the discretionary space from human decision making to the data wrangling process, where biases could still manifest. With autoethnography, the authors were able to critically examine their role as legal scholars navigating the tech field, revealing inherent complications and biases that arise when merging human judgement with ML tools.

Work from Faith, 2024 explores the author’s personal journey of reflection and intellectual development through extensive interactions with ChatGPT. Using an autoethnographic approach, the author narrates how initial utilitarian interactions with the system evolved into deeper cognitive and emotional explorations over time. Notably, autoethnography allowed the author to identify and report hidden biases and

behaviours that were previously overlooked, namely patterns of impatience, frustration, and high expectations. Similarly, King and Prasetyo, 2023 explore their experiences using generative AI for educational purposes and examine how their expectations and emotional responses shifted over time, providing insights for educators considering integrating AI in their teaching practices.

Based on the literature, self-study research methods seem to be particularly effective in exposing expectations, biases, and behaviours around AI. Unlike traditional user studies, which often focus on short-term interactions, self-study allows extended, genuine use of systems aimed at achieving real-world goals. This approach enables researchers to observe how cognitive processes evolve as they interact with AI in a more natural context. Capturing these patterns is important because ML applications are often integrated into daily workflows, where they influence decision making, user behaviour, and system performance in ways that become apparent only through sustained and goal-oriented use.

4.2 Motivation for Autoethnography

The purpose of this work is to extend the findings of the user study on TACA presented in Chapter 3. The user study revealed that participants, all non-experts in ML, were able to critically reflect on their qualitative analysis, adapt their interpretive stance, and gain new thematic insights through the use of the tool. The study also highlighted common misconceptions about ML concepts, as participants often viewed the model as an objective source of advice, attributing poor performance to their own analysis rather than to the system. Although the user study provided valuable insights into how non-experts interact with ML systems, there are several aspects of the study that could be addressed with autoethnography to gain a more comprehensive understanding of these interactions.

One aspect concerns the familiarity with the data that the participants used. IML applications generally assume a considerable degree of domain knowledge from the end user, as general familiarity with the data is required for accurate model inspection and feedback assignment (Amershi, Cakmak, et al., 2014). However, of

the 20 participants, only 5 imported transcripts from their own research, while the others received a collection of newspaper restaurant reviews to analyse beforehand (this compromise was necessary due to difficulties in participant recruitment, as eligible participants rarely had an analysed data set available or that met the structural requirements of the tool). As a result, those who used restaurant reviews may not have had the level of contextual understanding that could improve their ability to accurately inspect and provide feedback to the ML model. In contrast, I was highly familiar with my own data set and analysis, which was conducted shortly before this work, as opposed to some participants'. I also had access to all the original data, including codes and themes that were merged during the analysis, which I could revisit while using TACA.

Another reason for conducting an autoethnography is that participants in the user study interacted with the ML model relatively infrequently. Of the 20 participants, 12 re-trained the model at least once, and 5 re-trained it twice or more. This was largely due to their tendency, as non-experts, to overestimate the accuracy of the suggestions of the model. Moreover, the time the tool takes to re-train the model and load the tables usually takes a few minutes, which may have discouraged frequent re-training. Therefore, the insights gained from their experiences within the iterative IML cycle were relatively limited. Instead, autoethnography could provide more opportunities to explore the ways in which iterative feedback, trust, and critical thinking develop throughout the cycle by making use of the full potential of TACA.

An additional aspect to consider is the difference in motivation between the participants in the user study and my own as the researcher. Participants used TACA primarily to complete a task for the study and, in all cases, did not use the tool to truly improve their analysis, as they had no intention of utilising the insights they gained. As discussed in the literature, self-study allows extended and genuine use of systems aimed at achieving real-world goals to deeply understand the effects of the system on real practice (Neustaedter and Sengers, 2012). The use of TACA in this autoethnography is not only sustained but also genuine, because I was directly

invested in the insights gained through the analysis reported in this work.

Only non-experts in ML were eligible for recruitment in the user study, as one of the advantages of IML systems is that even users without technical expertise in ML can refine and improve the model through iterative feedback. In contrast, my experience with ML means that my behaviours and insights when interacting with the model are likely to differ significantly from those of the participants. For example, a study by Q. Yang, Suh, et al., 2018 revealed that non-experts are generally more satisfied and trusting toward the outcome of ML compared to their professional counterparts. This perspective is important because it allows for a valuable comparison between expert and non-expert user behaviour and insights into how to tailor ML applications to meet the needs of both groups.

Finally, autoethnography allows for the direct comparison between the results of performing a qualitative data analysis manually on the user study interviews and the results obtained when extending the analysis with TACA. This comparison provides a unique perspective, as I am able to evaluate my analysis before and after using the tool, focusing on how TACA influences the analytical processes. In contrast, participants in the user study inevitably spent much of their attention learning how the tool works, how to interact with its features, and how to interpret the results, which likely limited their ability to critically assess the analysis itself.

4.3 Reflexivity and Positionality

The positionality of the researcher arguably affects all aspects of their work (Fassl and Krombholz, 2023). This is especially true in autoethnography, which focuses on personal experiences. To provide the context needed to appropriately interpret the findings, this section outlines my educational background, prior experience, and positionality with respect to qualitative research and artificial intelligence.

My higher educational background began with a BSc in Computer Science which I chose after developing an interest in coding through a school project. My passion in computer science has always been tied to the rational structure of programming, as well as the reward of developing final products, whether algorithms

or applications. Finding HCI my favourite module of the course, I chose to pursue an MSc and a Ph.D. in the same field. In HCI, I found that I could both create applications and contribute to research by uncovering novel insights, particularly in how people interact with systems, how they navigate interfaces, and how technology shapes behaviours. A particularly rewarding moment occurred during my MSc final project, where I found statistical significance in a quantitative analysis of user behaviour around chatbots. My first exposure to working with AI was during this study. I found the experience particularly fascinating as it felt as though the system was operating beyond what I had explicitly coded, while simultaneously opening up unlimited opportunities for research into how people understand behave and interact with autonomous systems.

Validation from tangible results has always been the driving factor of my work. For this reason, transitioning from the logical and rational perspective of coding and quantitative research to qualitative research was challenging. The qualitative research process is subjective by nature (Braun and Clarke, 2019), and the need to obtain measurable results to validate my work was more difficult to meet. The shift from a somewhat positivist mindset to one that embraces complexity and ambiguity was necessary to broaden my understanding of human-AI interaction, and it occurred when I conducted a thematic analysis on interviews with participants in the study on TACA. During this process, I came to understand that conducting qualitative research is a systematic and rigorous process following several established guidelines and frameworks that can support the validity and reliability of the findings. Additionally, I realised that subjectivity does not invalidate the outcome, instead, it can often add depth and nuance to understanding human experiences. Although I believe I have moved beyond my initial biases, I recognise that I still approach qualitative research from an initial contrasting standpoint, shaped by my background in computer science and AI.

My interest in AI has grown significantly in the past few years as a result of the ongoing period of rapid progress in the field. I find it fascinating that algorithms can learn and adapt independently, achieving remarkable results that can sometimes sur-

pass human performance. I have incorporated AI applications in both my work and my life, using large language models to program more efficiently and answer general questions, often replacing traditional search engines. I have also experimented with image generation as a hobby and have learned more about machine learning theory for fun. Although I have first-hand experience of the limitations of AI and also recognise the great potential risks, my personal view of AI is undoubtedly positive. Admittedly, this perspective is heavily influenced by my own participation in AI research, where I am arguably less personally affected by concerns such as job displacement than those outside the field. However, I am convinced that AI can be used more as a tool to enhance human capabilities by improving efficiency, creativity, and problem solving in various industries rather than replacing human involvement altogether.

4.4 Method

4.4.1 Previous Qualitative Data Analysis Results

The autoethnography involves the use of a previously analysed data set: the semi-structured interview transcripts from the previous user study on understanding interaction with ML through TACA. These were transcribed verbatim from audio recordings then analysed using inductive thematic analysis. The content of the themes identified and how they contributed to the findings are discussed in this section. See Appendix B.1 for all the codes belonging to each theme.

4.4.1.1 Design Choices

This theme includes all comments from participants on the structural elements of the tool, such as the use of tables, colour schemes, and tooltips. These elements influenced how participants navigated the user interface of TACA and understood the output of the ML model. The theme includes discussions around the perceived and measured advantages and limitations of the design choices made when developing the tool. These insights contributed to the findings by showing how the UI significantly impacted the user's ability to engage with and understand ML models.

4.4.1.2 Data Review

This theme captures how participants approached reviewing their own analysis and the suggestions of the model within TACA. It includes how they explored patterns in the data, identified commonalities, and reflected on human error in qualitative data analysis, using the tool to better understand and re-evaluate their data. The findings showed that participants valued the ability of the tool to support reflection and re-evaluation of their own analysis. By reviewing their data, participants demonstrated increased self-awareness of coding decisions, especially when the model contradicted their initial interpretation.

4.4.1.3 Using TACA/Machine Learning in Qualitative Data Analysis

Here, participants reflected on the efficiency and limitations of using TACA and ML to support qualitative data analysis. They compared TACA to other tools, its purpose, use case, and potential to streamline the analysis process within their existing workflows. This theme contributed to the discussion by illustrating how participants appreciated the potential of ML to reduce manual labour of coding, including reflections on the need to complement rather than replace human interpretation.

4.4.1.4 Perception of the Model

Participants described how they interpreted the suggestions of the model and evaluated its performance. They often viewed the model as an objective source of advice, highlighting misunderstandings of ML concepts and a mental model mismatch. This theme contributed to the findings by exposing common misconceptions about the model, with participants overestimating its performance and demonstrating contradictions between their perceptions of the objectivity of the predictions and the subjective nature of the training data set.

4.4.1.5 User-Model Tensions

This theme concerns how participants experienced tensions in their interactions with the model, particularly around ambiguity in the data, class imbalances, and contradictions in the confusion tables. Additionally, it describes how the participants

justified the inaccurate suggestions of the model to their own mistakes during the coding process and how they trusted the model compared to their own abilities as researchers. The findings here revealed the participants' tendencies to blame themselves for poor model performance, even when errors stemmed from the limitations of the model.

4.4.1.6 Interacting with the Model

This theme is focused on specific interactions with the tables when re-labelling samples and re-training the model, mainly concerning dragging and dropping keywords or sentences across columns. It contributed by showing that participants appreciated the ease of interaction but often misunderstood the implications of keyword manipulation, as they believed that the predictions of the model were made on individual words rather than whole sentences.

4.4.2 Autoethnography Data Collection

The following plan of action was developed before conducting the autoethnography:

1. Revisit the coded semi-structured interview transcripts, the theme-code table, the results of the thematic analysis and the instructions of the study and the tool given to participants before the study.
2. Run TACA and explore keywords in the Codes Table.
3. Identify manual labelling mistakes in the Train Keywords Table.
4. Re-label and re-train the classifier until no perceived improvements in the Confusion Tables.
5. Switch to the Predict Keywords Table and identify classification mistakes.
6. Re-label and re-train the classifier until no perceived improvements in the Confusion Tables.
7. Switch to the All Keywords Table and explore relationships between key-words/sentences and themes.

8. Document usage and experience throughout using field notes, screenshots, and interaction logs.

The plan was significantly influenced by my previous experience working with ML for other aspects of my Ph.D. In particular, I knew from previous experience that the way text is labelled in thematic analysis initially does not work well for training a ML model. This is because coding units can range from very short phrases to entire paragraphs, which introduces ambiguity when converting coded data into structured labels for ML models. For example, I knew that TACA works by breaking the text into sentences before assigning a label to each sentence. However, if a code was assigned to an entire paragraph, it is likely that not every sentence actually belonged to that label, especially taken out of context. Therefore, I intended to approach the re-labelling process with a particular strategy that involved first focusing on correcting the training data set before attempting to re-label the classifications of the model.

This approach aimed to ensure that the data used to train the model was as accurate as possible, which I believed would lead to better predictions. I was convinced this was a solid approach as I had used a similar method to improve the performance of the same kind of ML classifier used for the study reported in Chapter 5. In that case, I manually reviewed the mistakes of the model using a spreadsheet, specifically instances where it incorrectly identified positive or negative cases (false positives and false negatives), and corrected any errors in the manual labelling. After several iterations of correcting the errors and re-training the model, I observed an improvement in the performance of the model, increasing the average F1 score (i.e., the F1 score averaged across all classes) from 0.60 to 0.84 on the test set.

Using Confusion Tables for the model inspection phase was also influenced by previous work on ML, but it was primarily based on the shared experiences of participants using TACA reported in the interviews. From the thematic analysis conducted on the semi-structured interviews, a key finding within the “Perception of the Model” theme was that participants could get meaningful insights about the state of the model from the frequency-based keywords in these tables. Participants

spent, on average, around 6:30 on the Confusion Tables, switching to them after re-training the model most of the times. Compared to the Keywords Tables, where participants used different strategies to assign feedback to the ML model, almost everyone focused on the false positive and false negative samples in the Confusion Tables to measure the performance of the model and identify areas for improvement. With this in mind, I adjusted the plan to include iteratively re-labelling samples in the Keywords Tables and evaluating the output of the Confusion Tables, specifically false positives and false negatives.

Despite the various strategies and approaches to analysing data in TACA due to personal experience with developing and interacting with ML applications, I still decided to follow the study and tool instructions as closely as possible. Adherence to the same guidelines would allow me to generate results that could be directly compared to those of the participants. For the same reason, the plan did not include making any changes addressing the shortcomings of TACA highlighted by the participants, such as reducing loading times. By choosing not to modify or improve the tool, my intention was to more accurately relate to the experiences of the participants and to provide a more authentic comparison of outcomes.

The study instructions specified a time limit of around 30 minutes for the use of TACA. However, a significant advantage of using autoethnography compared to the user study was that I could use the tool as long as necessary. This would allow me to engage with TACA through multiple iterations of model inspection and feedback assignment so that I could refine the model as long as it proved beneficial. The absence of a time constraint also meant that I could endure the relatively long loading times after each re-classification, which were noted by the participants in the post-study interviews. By disregarding this one instruction, my intention was to fully explore the iterative process and gain more meaningful insights into my experiences using TACA.

I decided to make use of field notes, screenshots and interaction logs throughout the process to document my experiences in real time. This decision was based mainly on the work on autobiographical design in HCI by Neustaedter and Sen-

gers, 2012, which highlights the value of documenting interactions with a system throughout genuine usage. A staple of autoethnographies, field notes provide a structured way to document not only actions taken, but also the emotions, challenges, and thoughts that arise during the process (Ellis, Tony E. Adams, and Bochner, 2011). Record keeping and data collection, including screenshots and interaction logs, would allow for a more comprehensive understanding of interaction by also capturing visual evidence and the quantitative aspects of the experience. Together, these methods were intended to create a rich and detailed narrative to reflect on the entire process.

The aim of this method was threefold. First, it allowed me to reflect on my usage of TACA during the different stages of importing the data set, identifying and re-labelling samples, re-classifying the model, and evaluating the model performance. This included collecting personal reflections on how the user interface and feedback mechanisms influenced my decision-making, labelling strategies and general workflow. Second, the method allowed for a deeper exploration of my personal perceptions, thoughts, and emotions during the interaction with the ML model, in the typical fashion of autoethnographies. Finally, a key objective of this method was to compare the results I obtained from extending the thematic analysis of the semi-structured interviews with the outcomes of the initial analysis conducted without the use of TACA.

4.5 Findings

In this section, I describe my experiences using TACA with the semi-structured post-study interview transcripts that were manually coded using inductive thematic analysis.

4.5.1 System Usage

My interaction with TACA took place over a span of 2 days as I frequently took breaks to annotate my experiences. During this time, I re-trained the model 7 times, moving 6 keywords and 51 sentences in the Train Keywords table, 4 keywords and 27 sentences in the Predict Keywords table, and 1 keyword and 11 sentences in the

All Keywords table. When inspecting the Keywords tables, I clicked on keywords to reveal the tooltip 97 times in the Train Keywords table, 75 times in the Predict Keywords table, and 33 times in the All Keywords table. Due to significant class imbalance, the initial average weighted F1 score across classes was low (0.25), but increased to 0.37 over the course of re-training.

4.5.2 Interacting with TACA

Designed to expose non-experts in ML to IML, TACA was developed over several months through an iterative process in which I regularly discussed features with my supervisors and quickly tested them to identify and resolve bugs or potentially challenges experiences that participants would encounter during the user study. Ensuring functionality was my priority during this time, which meant that there was little room to fully explore how the features of the tool could be used to meaningfully engage with the data. At the time, I also did not have a suitable data set from my own interviews and research that I could use to thoroughly test the functionality of TACA. With this awareness, I anticipated moments of frustration in which I would wish I had approached certain features or design choices differently as a designer and developer.

4.5.2.1 Class Imbalance

Following the plan drafted and described above, among my first interactions with TACA involved the exploration of the Codes Table (see Figure 4.1). Compared to the theme-code table as a spreadsheet I had revisited before running the tool, this table offered additional insights into the thematic analysis I had conducted manually, namely the number of sentences belonging to each code and theme (shown in parentheses) and the individual sentences belonging to each code.

My experience implementing, training and testing ML models in my work has taught me the importance of class balance in model performance, a key concept I also learned while studying ML theory. This particular knowledge influenced the attention I paid to the number of samples across different classes, in this case themes, as I understood that imbalances could negatively impact the accuracy and reliabil-

Codes					
design choices (886)	data review (1090)	using taca/ml in qda (718)	perception of the model (866)	user-model tensions (694)	interacting with the model (201)
confusion tables (355)	commonalities (51)	automated coding (48)	accuracy interpretation (124)	ambiguity (67)	change from one theme to another (80)
confusion tables limitations (17)	correct (129)	comparisons with other tools (45)	accurate suggestions (285)	class imbalance (21)	drag and drop interaction (33)
confusion tables terminology (26)	data exploration (156)	easier (39)	confusion between keywords as pointer	confusion tables contradictions (132)	keywords as handles (29)
confusion tables tooltip (7)	explorative approach (94)	efficiency of using ml (8)	confusion tables agreements (14)	data quantity (17)	losing information (7)
different colours (10)	human fallibility (65)	efficient (24)	confusion tables strategy (86)	different performance on different parts	re-classification improvement (3)
frequency (82)	information extraction (57)	experience in qualitative analysis (47)	data convergence (4)	expectations (106)	re-classification limitations (37)
frequency counters (40)	keywords context (109)	faster (14)	external source (126)	inaccuracy interpretation (119)	re-classification strategy (85)
general limitations (18)	keywords to explore results (180)	human-tool distinction (0)	higher level of ai (11)	inaccurate suggestions (110)	
improvements (116)	manual coding (213)	iterative process (40)	how the tool works (54)	justification of model inaccuracy (112)	
keywords (676)	manual coding process (150)	limited value of ml metrics (22)	improvement interpretation (3)	own lack of clarity (61)	
keywords tab (191)	own-coding feedback (259)	ml influence on analysis (65)	interpretation (163)	trust (72)	
keywords limitations (117)	perception of things being objectively cc	participant background (45)	keywords vs sentence embeddings (11)	trust in the model (19)	
keywords sorting (59)	reflexivity (146)	purpose (78)	lack of experience in ml (7)	trust in manual coding (30)	
keywords tables (187)	research question (25)	researcher vs tool distinction (61)	mental model (99)	user own fault (55)	
keywords tooltip (48)	results summarisation (109)	team collaboration (6)	missed sentences (222)	user-model comparison (179)	
limitations of other tools (37)	semantic relations (50)	theme identification (18)	model evaluation (111)		
meaningless keywords (70)	semantics (210)	theme adjustments (39)	objective (177)		
performance limitations (8)	strategic approach (25)	useful (253)	perceived limitations of the model (47)		
setup (16)	subjectivity (21)	use case scenario (62)	personalised model (45)		
sources of confusion (95)	supportive role (107)	user in control (8)	re-classification evaluation (76)		
text tab (57)		tedious (48)	transparency (0)		
			understanding (80)		

Figure 4.1: Screenshot of the Codes Table

ity of the suggestions of the model. Although TACA does handle class imbalance by implementing an oversampling algorithm for multi-label classification (Charte et al., 2015), the Codes Table revealed that the classes were highly imbalanced, leaving me somewhat concerned about the performance of the model. An initial strategy that I quickly discarded but that would stay in the back of my mind during the rest of the use of TACA was to favour re-labelling *towards* the theme “Interacting with the Model”, the theme with the lowest number of sentences, in an attempt to manually balance the classes. This consideration marked my first exposure to the tension between improving model performance and improving the analysis, a recurring theme in this autoethnography.

4.5.2.2 Data Set Structure

Clicking on individual terms in the Codes tables displays a tooltip containing the sentences that were manually labelled with that particular code. The sentences revealed in the tooltip confirmed my impression that generating a training data set for an ML model directly from qualitative data analysis is challenging and requires manual refinement to address the structural differences of the two types of data. In particular, I noticed that a considerable number of sentences, when taken out of context, did not actually belong to the code assigned. See Figure 4.2 for an example.

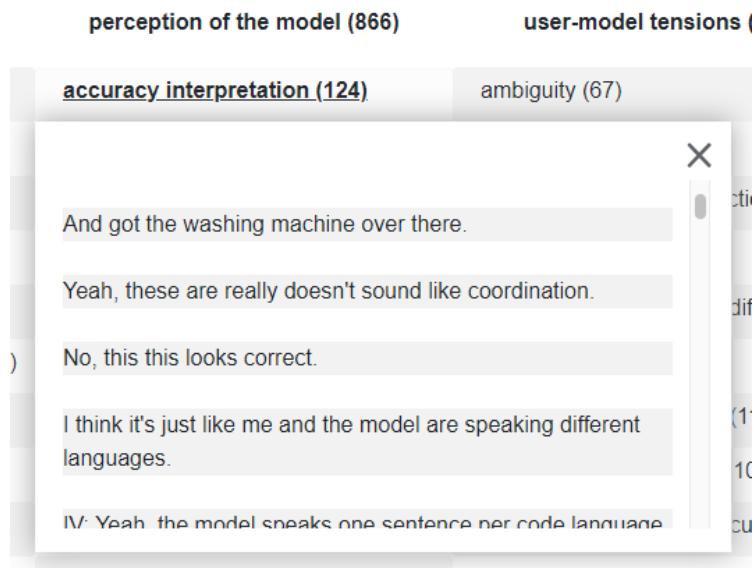


Figure 4.2: Example of a code including a sentence taken out of context: “And got the washing machine over there”.

These sentences are significant because the model is trained on the direct association between individual sentences and their assigned class label, without considering the broader context. For example, in the case of “And got the washing machine over there”, the model might incorrectly learn that “washing machine” is linked to the class “Perception of the Model”, leading to potential misclassifications. This sentence is part of a whole paragraph that was manually labelled as “Perception of the Model”. Here, a participant was reviewing sentences labelled by the model and (coincidentally) how they and the model “spoke different languages” when labelling text: contextually and verbosely compared to one sentence per code.

4.5.2.3 Re-labelling Training Samples

Part of the re-labelling strategy I chose to follow involved interacting with the Train Keywords Table first to correct the labelling mistakes that were either made during the coding phase of the thematic analysis or that appeared after converting a data set used for thematic analysis to one used to train an ML model. Shortly after starting to compare the meaning of each keyword with its respective theme to identify the ones that were out of place (the same way as participants in the user studies), I realised that there were simply too many keywords to go through (see Figure 4.3). As a

The screenshot shows a software interface for managing keywords. At the top, there are tabs: Text, Codes, Keywords (with a dropdown arrow), and Confusion Tables. Below this is a header row labeled "Train Keywords" with columns: "design choices (825)", "data review (1016)", "using taca/ml in qda (692)", "perception of the model (819)", "user-model tensions (650)", and "interacting with the model (191)". The main area is a 6x6 grid of cells, each containing a keyword and its frequency. The cells are color-coded in a light grey gradient. A tooltip "Re-classify" is visible above the first column. In the bottom right corner of the grid, there is a small window titled "Bin" with the text "Drag here to remove theme".

design choices (825)	data review (1016)	using taca/ml in qda (692)	perception of the model (819)	user-model tensions (650)	interacting with the model (191)
see (90)	see (103)	tool (68)	see (72)	see (53)	move (22)
keywords (78)	coded (50)	see (65)	theme (58)	theme (46)	keywords (22)
theme (39)	theme (50)	keywords (54)	keywords (58)	themes (37)	bin (16)
lot (35)	themes (48)	analysis (51)	tool (47)	coded (36)	see (15)
themes (34)	way (48)	coding (46)	coded (43)	false (31)	ones (13)
helpful (34)	keywords (47)	useful (43)	words (41)	keywords (30)	keyword (13)
quite (33)	lot (46)	code (40)	lot (40)	put (30)	theme (12)
words (33)	words (45)	themes (40)	false (38)	say (27)	word (11)
false (33)	coding (44)	qualitative (39)	missed (37)	coding (27)	moved (10)
coded (32)	say (43)	helpful (39)	say (33)	quite (26)	sentence (10)
word (31)	analysis (41)	use (38)	look (32)	ones (25)	put (9)
sure (27)	quite (39)	different (38)	positives (32)	lot (25)	sentences (8)
little (27)	interesting (39)	way (34)	themes (31)	code (24)	feit (8)
coding (27)	code (39)	coded (33)	way (30)	first (24)	codes (8)
positives (27)	tool (39)	lot (33)	time (30)	wrong (23)	drag (8)
time (26)	ones (37)	time (33)	codes (30)	helpful (23)	time (8)
look (25)	time (36)	theme (32)	different (30)	way (22)	another (7)
different (24)	different (36)	text (32)	good (30)	word (22)	words (7)
might (24)	might (35)	quite (32)	might (29)	words (21)	look (7)
ones (24)	back (35)	data (29)	code (28)	different (20)	want (7)
way (24)	helpful (34)	back (29)	ones (28)	good (20)	might (18)
confusion (24)	right (33)	might (27)	word (28)	right (19)	eve
put (23)	sentences (33)	good (27)	sentence (27)	might (18)	sur
useful (23)	word (33)	hello (26)	coding (26)	positives (18)	relevant (6)

Figure 4.3: Screenshot of the Train Keywords Table

result, I changed my approach by identifying only the one or two keywords with the strongest correlation to the respective theme in each column, and see whether those appeared in different columns too. These included, for example, the keywords “helpful” and “keywords” in “Design Choices”, “coded” in “Data Review”, “tool” in “Using TACA/ML in QDA”, “wrong” in “User-Model Tensions”, and “drag” in “Interacting with the Model”.

My interactions with the keywords evolved over time. Initially, I found myself sharing the same concern as the participants whenever I considered dragging and dropping keywords from one theme to another. Namely, that using keywords as handles for groups of sentences would result in re-labelling some sentences that I would not want to re-label. While interacting with the drag-and-drop functionality, I quickly realised that not all sentences including the most frequently appearing keywords should have been moved. Therefore, my earlier interactions involved moving individual sentences, mainly from the keywords with a strong association to a theme that did not appear under that theme. Realising that this process was not very efficient, I tried to identify keywords with a lower frequency number that were more likely to include only sentences that should have been moved, for example

Train Keywords					
design choices (825)	data review (1016)	using taca/ml in qda (692)	perception of the model (819)	user-model tensions (650)	interacting with the model (191)
keywords (76)	see (117)	tool (73)	see (66)	see (51)	bin (28)
see (73)	coded (83)	see (66)	theme (56)	theme (44)	keywords (28)
helpful (55)	keywords (56)	keywords (54)	keywords (55)	themes (35)	move (27)
theme (37)	look (56)	analysis (52)	coded (41)	false (32)	moved (20)
quite (36)	theme (55)	coding (46)	words (41)	coded (31)	see (17)
themes (33)	themes (52)	useful (43)	lot (41)	put (30)	keyword (17)
lot (33)	words (49)	code (40)	tool (40)	say (24)	theme (15)
words (30)	way (48)	themes (40)	false (37)	keywords (24)	ones (13)
word (27)	lot (48)	qualitative (40)	missed (37)	lot (24)	put (13)
coding (27)	coding (47)	helpful (38)	say (34)	ones (24)	word (13)
false (25)	say (46)	use (38)	look (32)	coding (23)	sentence (12)
way (25)	back (43)	different (38)	positives (32)	code (23)	drag (10)
sure (24)	interesting (42)	way (34)	way (31)	first (23)	sentences (10)
little (24)	analysis (41)	coded (33)	themes (29)	wrong (23)	felt (9)
time (24)	tool (41)	lot (33)	codes (29)	quite (22)	time (9)
confusion (24)	ones (41)	time (33)	good (29)	word (21)	codes (8)
put (23)	code (40)	theme (32)	different (29)	way (20)	words (8)
positives (22)	quite (40)	text (32)	code (29)	good (20)	want (8)
different (22)	false (40)	quite (32)	time (28)	words (19)	another (8)
ones (21)	different (39)	data (29)	might (28)	different (19)	around (8)
might (21)	time (38)	back (29)	ones (28)	right (19)	look (7)
useful (21)	helpful (38)	good (28)	word (27)	might (19)	sur (7)
first (21)	might (38)	might (27)	coding (26)	positives (18)	son (7)
		sentenna (75)	link (18)		check (7)

Figure 4.4: Screenshot of the Train Keywords Table after the initial re-labelling process.

“bin” (12 sentences) and “moved” (13 sentences) from “Perception of the Model” to “Interacting with the Model”, and “coded” (32 sentences) and “look” (24 sentences) from “Design Choices” to “Data Review”.

One aspect that emerged from this interaction with the Train Keywords Table is the concept of trade-off between the effort required (in terms of time and manual labour of re-labelling) and the quality of the re-labelling. In other words, I considered that it might be more efficient to drag and drop keywords containing a large number of sentences, even if not all of the sentences should technically be re-labelled, as the net effect on steering the model could still be positive. However, since time was not necessarily an issue, and in fact autoethnography allowed for full use of TACA, I decided to commit to a more fine-grained re-labelling approach. See Figure 4.4 for the Train Keywords Table after the changes made to address the mistakes due to manual labelling and the sentence extraction process.

4.5.2.4 Using Confusion Tables to Drive Model Feedback

Confusion Tables were designed to drive the model inspection phase in the IML cycle. Inspired by confusion matrices, the intention was to enable users to see where the model was performing well and where it was not, potentially suggesting which

samples to re-label before re-training the model again. Here, my habits of working with performance metrics were difficult to break, and I found myself adding the number of true positive and true negative sentences, comparing it to the number of false positives and false negatives. At this point, I was convinced that improving the performance of the model was not the priority as the researcher and participant of the autoethnography. However, gauging which theme required the most re-labelling was still something I thought to be useful in terms of efficiency.

At a first glance, my initial concerns about class imbalance were somewhat confirmed. The theme “Interacting with the Model”, which contained the lowest number of sentences in the training data set, was heavily skewed towards negatives in the Confusion Table (see Figure 4.5). Regardless, I still approached the tables with the goal of reviewing mainly the false positive and false negative classifications. This strategy was informed by previous experience from other work in my Ph.D., but also by the participants in the user study, who demonstrated the most critical and analytical thinking when reviewing these specific columns in the table. Perhaps a clear difference in my approach was the additional step of identifying the sentences where the classification of the model looked more convincing than my own labelling, and consequently re-labelling similar sentences in the Keywords Tables.

An example that illustrates the process of identifying sentences appearing in the Keywords Tables that were similar to the ones appearing in the Confusion Table involves the keyword “move”, or “moved”. Together, these were the keywords that appeared the most frequently under false negatives for the theme “Interacting with the Model”. Looking for the words in the Keywords Tables proved to be tedious, as there were definitely too many words for me to quickly find them. Here, I wished I had implemented a way to search individual terms, something that I recalled one participant mentioning during the interviews. Dragging and dropping “move” in the Train Keywords Table from “Using TACA/ML in QDA” (7 sentences) to “Interacting with the Model”, I truly felt like I was engaging with TACA according to the iterative cycle discussed in the literature on IML by giving feedback to the model

Text	Codes	Keywords ▾	Confusion Tables ▾
"Interacting With The Model" Confusion Table			
True Positives (7)	False Positives (10)	True Negatives (393)	False Negatives (48)
bin (3)	word (2)	see (29)	moved (6)
move (2)	classified (2)	keywords (27)	keywords (5)
gonbin (1)	themes (2)	lot (23)	move (4)
umim (1)	told (1)	theme (18)	sentence (4)
initially (1)	training (1)	code (17)	keyword (3)
tended (1)	boss (1)	tool (16)	time (3)
around (1)	realized (1)	helpful (16)	place (3)
interesting (1)	role (1)	different (16)	different (3)
playing (1)	ai (1)	false (15)	word (3)
check (1)	definitely (1)	ones (15)	thinking (3)
put (1)	helpful (1)	first (15)	codes (3)
better (1)	precise (1)	useful (14)	sure (3)
inconvenience (1)	review (1)	themes (14)	sentences (2)
felt (1)	checking (1)	way (14)	check (2)
sentences (1)	every (1)	say (14)	felt (2)
moved (1)	trying (1)	words (14)	made (2)
still (1)	keep (1)	good (13)	first (2)
sentence (1)	check (1)	quite (13)	coded (2)
click (1)	time (1)	text (13)	theme (2)
look (1)	changing (1)	put (13)	confusion (2)
words (1)	either (1)	might (13)	nothing (2)
removed (1)	theme (1)	analysis (13)	sense (2)
word (1)	usually (1)	true (13)	wrong (2)
friend (1)	belong (1)	codina (12)	terms (2)

Figure 4.5: Screenshot of the Confusion Table for the theme “Interacting with the Model”.

according to the model inspection phase.

A limitation of this strategy that I considered and that restrained me from making more than a few movements according to the false negative keywords is that the Keywords Tables do not include sentences that are unlabelled. The false negative samples in the Confusion Table could also have been classifications in which the model did not assign *any* label. However, the Keywords Tables only include sentences that have been assigned at least *one* label. Therefore, there were simply fewer words or sentences appearing in the Keywords Tables that could be similar to the ones misclassified as false negatives. After my iterative interactions with the Confusion and Keywords Tables, I wished I had implemented a feature that would allow users to accept or decline changes directly from the Confusion Table to avoid switching tabs, finding the right keywords, and also address the limitation mentioned above.

4.5.3 Perception of the Machine Learning Model

4.5.3.1 Initial Considerations on Model Performance

I experienced conflicting views about the ML model used in TACA when I began the autoethnography. My Ph.D. experience taught me that tuning hyperparameters

is crucial for achieving good performance in ML. Typically, hyperparameters set to general-purpose values may result in subpar performance for specific data sets or tasks. However, I was aware of a key decision made during TACA’s development, namely to avoid optimisation and use default values instead. This choice was made for several reasons. Hyperparameter optimisation would need to occur for every user, as the process is highly dependent on the data set used. This computationally intensive process requires training and testing the model multiple times on different data splits, significantly increasing initial loading times. Additionally, TACA is an IML tool designed to investigate the process of re-labelling and re-training by allowing users to iteratively improve the model through interaction. In this context, some initial model inaccuracy was actually beneficial.

The assumption made when deciding to avoid optimising the parameters of the model was therefore that users would be able to improve its performance through the process of re-labelling and re-training. However, this was not actually confirmed during the user study, as the participants did not re-train the model enough times to determine whether the performance was genuinely improving. Instead, my hopes for model performance were based on the study on interpretability techniques reported in Chapter 5, in which I significantly improved the same model by just correcting the ground truth by re-labelling samples in the training data set.

Regardless, it was clear from the interviews in the user study that the participants demonstrated thoughtful reflection on their coding and gained valuable insights from the data from suggestions that they considered to be “*wrong*”. For the purpose of exploring a qualitative data set, perhaps achieving good model performance was not as important as I had originally thought.

4.5.3.2 Reviewing Suggestions

The very first consideration I made when reviewing the suggestions generated by the model was about my own selective bias involved in the process. In the absence of performance metrics, the only way to evaluate the model is through individual classifications. As an invested developer, I found myself giving more weight to the samples that I thought were correctly classified while downplaying or rationalising

Predict Keywords					
design choices (102)	data review (246)	using taca/ml in qda (68)	perception of the model (79)	user-model tensions (40)	interacting with the model (53)
keywords (22)	see (29)	analysis (13)	see (13)	see (7)	move (10)
see (10)	word (13)	use (10)	false (10)	sentence (4)	keywords (9)
true (6)	look (12)	qualitative (9)	theme (8)	false (4)	theme (8)
sometimes (5)	way (11)	coding (9)	words (8)	let (4)	sometimes (6)
false (5)	theme (11)	see (9)	say (7)	sometimes (3)	drag (6)
themes (5)	analysis (11)	theme (8)	sentence (6)	codes (3)	see (6)
helpful (5)	sort (11)	tool (8)	code (6)	themes (3)	moved (5)
used (5)	say (10)	coded (7)	use (6)	positives (3)	sentence (4)
words (5)	words (9)	project (6)	positives (6)	way (2)	around (4)
table (5)	lot (8)	keywords (6)	themes (5)	lot (2)	bin (4)
theme (5)	want (8)	helpful (5)	example (5)	true (2)	felt (3)
positives (5)	themes (8)	text (5)	keywords (5)	transcript (2)	keyword (3)
people (4)	code (8)	codes (5)	different (5)	using (2)	reclassify (3)
sentence (4)	keywords (8)	quite (5)	lot (5)	correct (2)	ones (3)
say (4)	said (7)	computer (5)	even (5)	still (2)	little (3)
sentences (4)	looking (7)	data (5)	accurate (4)	word (2)	another (3)
quite (4)	sure (7)	fit (5)	still (4)	sort (2)	fit (3)
still (4)	coding (7)	might (4)	looking (4)	sure (2)	saw (3)
bin (4)	codes (7)	themes (4)	codes (4)	ones (2)	let (3)
way (4)	example (7)	sort (4)	correct (4)	getting (2)	want (2)
tables (4)	ones (7)	useful (4)	model (4)	question (2)	single (7)
confusion (4)	restaurant (7)	lot (4)	used (4)	wrong (2)	cod
little (3)	place (7)	looking (4)	another (4)	coding (2)	two
want (3)	out (6)	table (4)	terms (4)	aesthetics (2)	themes (2)

Figure 4.6: Screenshot of the Predict Keywords Table after the initial re-labelling and re-training process.

the mistakes. This was possible due to the absence of a clear ground truth, allowing me to navigate the ambiguity of the sentences and interpret the classifications in a way that fit my view of an accurate ML model. However, as the researcher in the autoethnography, I needed to resist the urge to justify the classifications simply because the model was the product of my work. Otherwise, I would not have interacted with the system as much as I should have to gain deeper insights from the data.

After several iterations of reviewing the false negative and false positive samples in the Confusion Table and re-labelling the related samples in the Train Keywords Table, the Predict Keywords Table looked surprisingly good despite the selection bias I had recognised (see Figure 4.6). In fact, I believed the most frequent keywords appearing at the top of each column had a clear connection with each theme, at least according to my manual coding process. My perception of the ML model improved significantly at this point, as I began to give more value to the suggestions.

Across every theme, I noticed that the coding suggestions of the model contained sentences that I had missed in the manual coding process. At this point,

I began to consider the model as a true extension of my own analytical process. Compared to the participants in the user study, who found the model to be an objective source of advice, I recognised the suggestions more as the product of the collaboration between my own interpretations and the ability of the model which shared my perspective. In fairness, this was probably because I did not notice many examples of suggestions that contradicted my own interpretation. Still, my trust in the model evolved over time as I perceived that it was genuinely improving.

4.5.3.3 Inner Workings of the Model

When reviewing the suggestions of the model, I realised that I was often trying to understand the reason behind the classification to improve my understanding of the current state of the model and inform re-labelling. The absence of interpretability techniques, such as word importance heat maps, meant that I was relying solely on speculation. However, my knowledge about how the model worked was enough to at least give me hints. For example, I knew that each word carried more weight to the classification in a short sentence compared to a long sentence. This is because the sentence is converted into an embedding vector, which is the average of all the embedding vectors for each word. With this in mind, I paid closer attention to short sentences containing important words, both when evaluating the performance of the model and when choosing which sentences to re-label.

Another aspect that I considered was the probabilistic nature of the model. To each classification, there is an associated confidence score that is not displayed in TACA. I wished I had access to the scores, as I would have directed my efforts toward re-labelling samples which the model classified confidently, because those are the instances that, when re-labelled, would have a greater effect on model behaviour. Compared to word importance and sentence length, I believed there was nothing I could use to compensate for the lack of metrics and cues. The design decisions originally made to facilitate the interaction for non-experts in ML made it sometimes frustrating to work with TACA as my workflow felt somewhat limited by the tool.

4.5.4 Emerging Insights and Analytical Reflections

4.5.4.1 Additional Insights

The initial process of reviewing the Train keywords with the strongest correlation to each of the themes already provided significant insights into the data that I had not noticed during the manual thematic analysis. For example, I believed “keywords” was one of the keywords most strongly associated with “Design Choices” and “Data Review” and, in fact, it appeared frequently in both columns. Keywords were arguably the most important design choice when developing TACA, as they were the primary means for the user to influence the behaviour of the model, and were also a means to facilitate data inspection by aggregating large amounts of data. However, “keywords” and “words” also very frequently appeared in “Perception of the Model”. When reviewing the list of sentences, it was clear that participants also experienced the model through the keywords, for example:

“The Predict keywords are based on your style of categorising the codes and generating the themes; it follows your patterns and tries to do your work as best it can.”

Although I did not find the fact that participants shaped their perception of the tool primarily through keywords particularly surprising, it was still something that I had initially missed. In fact, the theme-code table did not include codes related to keywords under the theme “Perception of the Model” except for one related to the confusion between pointers and semantics. This is important because it would have been valuable to explore how keywords, as a design element, impacted the perception of the model, especially compared to more traditional evaluation measures, such as performance metrics.

Another example of a different insight that emerged from just reviewing the data in the Train Keywords table concerned the distinction that participants missed between the tool and the model. In particular, I noticed that the word “tool” appeared very frequently in the theme “Perception of the Model”. After inspecting the sentences, it was clear that many participants referred to the ML model as just “the tool”, for example:

“and there was a direct quote that they had inside of the review, and then the tool chose something next to it that looked similar.”

Looking back at the interview transcript, this participant never once used the term “model”, even when answering questions that specifically mentioned it. In the original Train Keywords table, “model” only appears 20 times in “Perception of the Model”, 13 times in “User-Model Tensions”, 13 times in “Data Review”, 12 times in “Using TACA/ML in QDA”, 11 times in “Design Choices”.

Coding suggestions from the model also provided several unique insights into the data. Besides revealing the sentences that I had simply missed during the coding process, the suggestions also included interesting patterns. For example, the keyword “true” appeared frequently in the Predict Keywords Table under “Data Review”. The sentences that included this term were all about “true positives”. Conversely, “false” appeared as a frequent keyword under “User-Model Tensions”. The manual data analysis already revealed that false positive and false negative classifications created some tensions with the model. However, it had missed that true positives played a different but equally significant role in reviewing manual coding.

4.5.4.2 Structural Changes to the Analysis

Some participants in the user study reported that, had they used TACA at an earlier stage in the analysis, their results would have been different. For example, they might have changed or added themes based on the predictions of the model in the Confusion Tables under false positive and false negative. With this in mind, I approached the suggestions of the model ready to question the results of my own analysis.

Although I initially believed that my analysis did not require a complete thematic restructuring after interacting with TACA, further reflection led me to consider alternative interpretations and some necessary adjustments. These were mainly driven by the fact that the same keywords appeared in multiple themes. For example, in the All Keywords Table, “keywords” appeared very frequently in every theme. In almost every case, I thought the predictions of the model made sense, and that the term “keywords” was simply used in different contexts. This made me con-

sider the possibility of having “keywords” as a separate dedicated theme. However, doing so would also require re-evaluating other themes where “keywords” previously played an important role, such as “Design Choices”, as it would disrupt their overall balance and coherence. In particular, isolating keywords into a dedicated theme might obscure the interplay between user interaction and model perception.

Another keyword that appeared very frequently in every theme was the term “see”. Initially, I had not given this term much thought, as I considered it simply a stop word that I had forgotten to filter during the setup of TACA. However, upon closer inspection of the sentences that contained the word, it was clear that “see” was used either as a synonym for “check”, or as a reference to the visual interface of TACA. In “Data Review”, these sentences were understandably about checking or verifying the data. Meanwhile, in “Interacting with the Model”, they were mostly referring to specific elements of the interface of the tool. The distinction between inspection-based interactions and visual elements was generally very subtle:

“When I click the keyword and see the sentences, I think it’s probably more linking to the overall restaurant environment.”

I soon realised that the weak distinction was not just about the use of the word “see”, but extended to the two themes “Data Review” and “Interacting with the Model”. Put simply, participants often reviewed the data as a natural part of their interactions with the model.

Initially, “Interacting with the Model” was meant to focus specifically on the drag-and-drop re-labelling within the UI. However, during the analysis, it expanded to include other aspects beyond the interface, such as the code “re-classification improvement” (see Appendix B.1). After my own use of TACA, I realised that these two aspects should have been separated. Similarly, the theme “Data Review” could have been split between reviewing the participant’s own manually labelled data and reviewing the suggestions of the model. These distinctions would have provided clearer insight into the in reviewing different types of data and the specific interactive qualities of the tool.

4.5.4.3 Reflexivity

Like the participants in the user study, engaging with the predictions of the model allowed me to critically reflect on my own presence during the qualitative data analysis process. One key difference, however, was that I did not consider false positive and false negative classifications as contrasting suggestions. Participants considered these to belong to an equally valid, if not better, interpretation of the data, and recognised their own role in the analysis when justifying their own labelling against them. In my case, I did not give as much weight to the classifications that contradicted my labelling and, instead, used them to drive model feedback by showing me on which samples the re-labelling should take place. Therefore, my experience of reflexivity was not specifically driven by tensions with the model, but by how the suggestions reinforced my existing decisions while also prompting me to consider alternative interpretations.

The structural changes to the analysis that I considered and reported above were mainly the product of interacting with the All Keywords Table. Necessarily, I had to compare these changes to my own manual analysis to evaluate whether the insights I gained warranted hypothetical adjustments to my theme and coding structure. For example, I came to the conclusion that adding a dedicated theme for keywords would contrast my analytical perspective, since it would lead to changes in the other themes that would place more emphasis on specific features of TACA. I recognised that the user study was run to understand how non-experts interact with IML, and that the aim of my role in the analysis was to uncover findings that could be generalised beyond TACA.

The awareness of my position also influenced my assessment of the second structural change. In this case, I realised that splitting “Interacting with the Model” into separate themes, one focused on the specific UI interactions and another on broader interactions with the suggestions, could have been beneficial. This realisation was partly driven by my recognition of bias as the developer of TACA. Looking at the data, I noticed that I had been viewing interaction with the tool primarily through the features I had implemented and consequently grouped all other interac-

tions under this single theme.

4.6 Discussion

4.6.1 Reflecting on the Different Roles in the Autoethnography

Autoethnography is so far an under-explored approach within the context of human-AI interaction and IML. The following section discusses the tensions of adopting the roles of a developer, a researcher, and a participant in using TACA.

4.6.1.1 Developer-Researcher Tensions

As the developer of TACA, I was directly invested in the outcome of this autoethnography. To some extent, my expectations of the tool were shaped by the user study, which exposed me to the participants' feedback regarding the benefits and limitations of TACA in supporting qualitative data analysis. Still, part of me also recognised that the participants were non-experts in ML, and that my professional background and in-depth knowledge of the inner workings of tool could uncover findings that would cast TACA in a more favourable light. Of course, an external researcher would not have approached the autoethnography with the same expectations and inherent biases.

During the autoethnography, there were moments in which the two roles were somewhat conflicting. For example, I realised that I was giving more weight to samples that I thought were correctly classified compared to those that were misclassified. I made a conscious effort to analyse both correct and incorrect classifications equally, regardless of my initial judgments, to engage with the model in more depth. This was possible by reminding myself that the performance of the model was not crucial to the autoethnography, and that the focus of this work was instead on my interactions with TACA.

Previous work has identified the tensions between the role of developer and researcher in autobiographical design, which draws on extensive, genuine usage by those creating or building the system. In an attempt to address “the complexities of using this method with more precision and finesse”, Riordan, 2014 and Desjardins and Ball, 2018 recommend sincerity, defined as “concerned with the degree

to which a study is marked by honesty and transparency”. To my best ability, I have reported my perceived biases and the limitations of the tool, as well as describing my positionality as a researcher to contextualise the findings. Situating the findings this way required significant introspective effort that I hope will inspire future researchers when considering their own positionality while engaging in autobiographical design and autoethnographic studies at the intersection of HCI and AI and beyond.

4.6.1.2 Developer-Participant Tensions

The participants in the user study were non-experts in ML. In comparison, not only did I have experience in ML, but I also knew exactly how TACA worked. Throughout the autoethnography, I employed strategies to re-label samples and inspect the model that were uniquely based on my experience and knowledge. For example, I knew that the sentence embeddings were generated by taking the average embedding of each word, and therefore preferred re-labelling shorter sentences that contained important words. These insights were simply not possible to gain from users exposed to the tool through the interface only, which did not reveal the underlying processes or provide explanations for how the model made decisions.

The effects of the role of the developer during the autoethnography were also felt in other interactions. Much of the effort in developing ML systems is made in achieving good model performance, and my own experience reflected that. I was also expecting that, initially, the performance of the model was not going to be high for a variety of reasons related to the training data set generation and hyperparameters selection. Together, these factors contributed to the tension between wanting to “fix” the system and remaining immersed in the analysis.

The perspective of the developer when interacting with TACA introduced an additional component compared to the previous results of the user study with non-expert participants. This perspective revealed different but arguably equally valid findings. In fact, although it is true that one of the advantages of IML is that model refinement can be driven by non-experts in ML, these applications are often used by experts and developers too due to the speed of the iterations and immediate

feedback (Amershi, Cakmak, et al., 2014). Therefore, reporting on how a user with experience in ML engages with TACA contributes to a more holistic understanding of IML by complementing the findings of the user study.

4.6.1.3 Researcher-Participant Tensions

The tensions between the roles of researcher and participant are widely recognised as part of self-study research methods. The priority of the researcher is to objectively analyse the behaviour of the participants, but as a participant myself, my subjective experience and involvement complicated that analysis. As a result, I have attempted to address the three measures of autoethnography according to the literature (Ellis, Tony E. Adams, and Bochner, 2011): reliability, validity and generalisability.

To answer the question of reliability, I have included factual evidence, where relevant, with screenshots of the different views of TACA as well as direct quotes from the data set. The decision to omit the timestamped interaction logs was made after considering that the use of the tool spanned several days, with many occasions in which I would stop interacting with the tool to take notes.

Validity refers to the extent to which the story enables the reader to enter the subjective world of the teller (Lucero, 2018). In an attempt to evoke in the readers a feeling of a lifelike, believable, and possible experience, I have described my thought processes, emotions, and biases, and how these were shaped by my personal background (Ellis, Tony E. Adams, and Bochner, 2011). Hopefully, even if the reader did not necessarily share my perspectives, they were still be able to relate to my experience.

In the absence of a participant pool, the generalisability of autoethnography moves from respondents to readers and is determined by whether the autoethnographer is able to illuminate unfamiliar processes (Ellis, Tony E. Adams, and Bochner, 2011). During this study, I was aware that specific elements of the inner workings of TACA would be unfamiliar to the reader, and paid close attention to describing them in detail. Additionally, as IML is currently a niche approach to ML compared to traditional systems, I also tried to describe the thought processes behind

key decisions made in the iterative cycle.

4.6.2 TACA Effectively Supports the Refinement of Qualitative Data Analysis

One of the objectives of this work was to directly compare the results of a qualitative data analysis performed manually on the user study interviews with the results obtained when using TACA. In line with my plan of action, I engaged with the tool through multiple iterations of model inspection and feedback assignment as long as it proved beneficial in terms of model performance and insights gained.

4.6.2.1 Findings Extending Previous Results

The integration of ML in the analysis revealed that I had missed numerous sentences during the coding process. In some cases, these oversights were inconsequential, as they would not have drastically changed the results of the analysis. In others, I became aware of patterns that the model had learned that I had not. For example, placing sentences including the keyword “false” in “User-Model Tensions” was expected, but the keyword “true” in “Data Review” revealed that participants used the true positive classifications to analyse their own coding. This finding would have been interesting to report, as it would have introduced an additional use case for the Confusion Tables that was unexpected during the design process, since these were implemented to drive model inspection.

More significant extensions to the results of the analysis were driven by reviewing the frequency of keywords across themes in the training data set, as well as the suggested sentences. For example, the conflation of identities between the model and the tool was missed without these frequency-based tables, which highlight patterns that are not always immediately apparent after the coding process. This aspect extends the original analysis by uncovering subtle interactions between users and TACA, suggesting that non-experts may fail to distinguish between the interface and the underlying algorithm. Without a clear understanding of the separate role of the algorithm, participants were more likely to trust the output of the model uncritically. Because the participants experienced the model and the tool as

a single entity, their experiences with the interface of the tool likely influenced their perception of the ML model. This finding, as well as explaining the perception of an impartial and objective model, has important implications for how to design and communicate ML-driven system to users, especially in an attempt to foster transparency and trust in non-experts.

4.6.2.2 Findings Contradicting Previous Results

Reviewing the All Keywords Tables, which included both the training and predicted data sets, revealed possible changes to the structure of the themes and codes that I had originally drafted. In particular, the fact that certain keywords, which were not stop words, appeared in every theme made me realise that perhaps the categories were not so distinct after all. Thematic analysis is not a rigid process, and themes can be flexible to accommodate the complexity of the data without forcing predefined categories (Braun and Clarke, 2006). However, I believe that restructuring the themes into more fine-grained categories would have led to a more nuanced analysis of the data.

An important structural change to the analysis that I believed would have been beneficial is the one related to the distinction between interacting with the model and interacting with the UI. When reporting the results of the user study, the two types were combined, but after reviewing the sentences in TACA, it was clear that separating these interactions would have allowed for a more precise analysis of how participants navigated the interface used in re-labelling versus how interpreted the re-classifications of the model. In fact, what emerged through my use of TACA is that participants reflected on their interaction with the drag-and-drop feature, revealing misconceptions around using keywords as handles for sentences versus keywords as features for the classification. They also described their interactions with the model overall, such as their re-labelling strategies and the perceived improvements in the outcome. These two interactions were related, but arguably did not belong to the same theme, as one required more mechanical actions, while the other required deeper cognitive engagement with the data. Distinguishing between these types would have highlighted the different processes at play and provided clearer

implications of the findings. For example, by separating mechanical and cognitive tasks, system designers can create more tailored interfaces to help users better understand the distinct roles of the interface and the algorithm.

Another unexpected discovery involved the term “see”. Initially dismissed as a stop word, it emerged as a significant term upon closer inspection of its usage. Participants used this term in two distinct ways: as a synonym for “check” during data verification in the “Data Review” theme, and as a reference to interacting with the visual interface of TACA in “Interacting with the Model”. This distinction prompted me to rethink how these themes related to each other. The subtle overlap between reviewing data and interacting with the model was not just about terminology but about the processes participants engaged in while navigating the interface. In particular, participants’ engagement with the model was tightly coupled with reviewing data, blurring the lines between these two themes.

This reflection pushed me to reconsider whether my initial structure was appropriate to capture the nuanced user behaviours revealed by TACA. Although I initially resisted overhauling the thematic structure, the iterative engagements with the tool allowed me to potentially refine the analysis by making distinctions between themes that would capture more fine-grained interpretations of the participants’ interactions. This divergence from my initial findings highlights the value of the iterative process in qualitative analysis, where revisiting data with new perspectives can reveal complexities and contradictions that were previously missed.

4.6.2.3 Encouraging Reflexivity in the Analytical Process

In addition to refining the structure of the analysis, TACA also encouraged reflexivity, one of the pillars of qualitative data analysis (Fontana, 2004; Jootun, McGhee, and Marland, 2009; Braun and Clarke, 2019; Holmes, 2020). Like the participants in the user study, I was frequently reminded of my role in the analysis. However, this happened for participants whenever they were justifying their own labelling when confronted with the contrasting classifications of the model in the false positive and false negative samples. In my case, I noticed the effects of my role when faced with potential changes to the themes after reviewing the entirety of the data. This finding

is aligned with the fact that users with more experience in ML are generally less trusting of the output of the model and, therefore, when the output is contrasting, it is less likely to be viewed as a valid alternative interpretation of the data (Q. Yang, Suh, et al., 2018). Designers of ML-driven systems should carefully take into consideration the users' expertise and previous knowledge in ML to design appropriate interactions and feedback mechanisms.

Ultimately, TACA helped me identify parts of the text that I had originally missed, evaluate areas for potential restructuring of the analysis, and encouraged me to reflect on my own positionality as the researcher. These insights enhanced both the breadth and depth of the qualitative analysis, demonstrating how TACA can effectively support qualitative data analysis and, more broadly, how ML can be integrated into analytical processes.

4.6.3 Implications of Design Choices on Interactive Machine Learning

The interaction with TACA revealed several aspects of the design choices made during development that have relevant implications on the interaction with AI-driven systems. Some of the considerations made during the autoethnography were shared by the participants in the user study, while others are in contrast to the behaviour of the participants due to the difference in expertise in ML concepts.

4.6.3.1 Batch Re-labelling and its Impact on Model Refinement

Batch re-labelling of data samples was a feature of TACA that was originally implemented to address an important limitation of IML: not only is labelling data tedious and sometimes not considered worthwhile by the user, but it requires investing significant effort before noticeable change in the model (Wong et al., 2011; Groce et al., 2014; Ribeiro, Singh, and Guestrin, 2016). The implementation of keywords was meant to exploit the most frequent terms as handles for groups of sentences, enabling the user to re-label multiple data points, or sentences, simultaneously.

The participants in the user study approached the keywords with mixed strategies: some were cautious, carefully considering the impact of re-labelling multiple

sentences at once, while others were more willing to experiment and quickly re-label large batches. My own experience with keywords fell into the first group, at least initially. The absence of a time limit in the autoethnography meant that I could adopt a more fine-grained approach to re-labelling, since the effort taken to re-label individual sentences was personally not an issue. However, I did begin to consider the trade-off between effort required and the quality of the re-labelling as the number of data points re-labelled increased.

For users who are constrained by time or resources, batch re-labelling might offer a more practical solution, even at the cost of potentially introducing errors in the data set. This is because re-labelling multiple data points at once can enable users to make faster progress and see immediate changes in the performance of the model. Conversely, users who can afford a more meticulous approach can improve the performance of the model by carefully refining individual data points and potentially achieving greater performance at the cost of more manual effort.

Interfaces for feedback assignment in IML require the most careful design in terms of both elements and interaction methods (Dudley and Kristensson, 2018). In an attempt to address the issues of batch re-labelling, visualising anticipated changes could introduce transparency in the system, which greatly affects the quality of the response elicited from users (Amershi, Cakmak, et al., 2014). However, in general, the design of IML systems should account for different user needs and offer flexibility for those who prioritise efficiency and accuracy differently based on their constraints and goals.

4.6.3.2 The Role of Transparency in Interactive Machine Learning

An aspect of my experience with TACA that differed from the participants in the user study is my preference for greater diagnostic transparency within the system. The results of the user study supported the need for interpretability techniques designed for model diagnostics to help users understand why the ML model presented a conflicting coding suggestion and avoid the perception of an objective source. However, no participant explicitly stated the need for deeper insights into the internal workings of the model or requested more detailed diagnostic information

beyond understanding the immediate outputs. In contrast, I became aware that transparency would facilitate feedback assignment by exposing the samples that, when re-labelled, would steer the ML model more efficiently toward more accurate predictions.

Interpretability is usually discussed in the literature in terms of how it positively impacts trust, reliability, robustness, causality, and usability of ML models (Doshi-Velez and B. Kim, 2017), particularly in high-stakes areas such as health-care, finance, and criminal justice (Samek, Wiegand, and Müller, 2017). Examples of implementations in existing tools include: model-specific, gradient-based techniques for Convolutional Neural Networks used for clinical diagnosis in medical image analysis, inherently interpretable models that support rule extraction for explanations and dialogues in conversational systems to predict legal outcomes in the justice system, and rule extraction methods used on interpretable models such as decision trees in financial services (Ding, Abdel-Basset, and Hawash, 2022; Srinivasu et al., 2022).

Compared to these applications, where model transparency is used to extract diagnoses and explanations from the model, interpretability could take on an alternative role in a tool like TACA. TACA encourages users to reflect on the data (both training and predicted) and build appropriate strategies to steer the model in a particular direction. In this context, interpretability can optimise the user's contribution to the iterative learning process of the model. Recent work by Teso et al., 2023 confirms that techniques can significantly enhance the user's ability to correct the model by providing explanations that highlight misclassifications and areas of improvement. They also note key challenges, such as ensuring explanations are both understandable and actionable, especially for non-experts, without overwhelming the users with too much information, which was the primary concern when developing TACA.

In general, transparency can also greatly reduce the distance between the experiences of non-expert participants reported in the user study in Chapter 3 and those detailed in this autoethnography. A notable difference was that the participants

were not aware of misclassifications due to the model making mistakes. In this case, word importance heat maps could visually highlight which words contribute most to the classifications of the model, making it easier for non-experts to identify when the decisions of the model are based on misleading terms rather than the broader meaning of the sentence. Alternatively, the presence of classification scores alongside model predictions could offer another form of transparency by indicating the confidence of the model in each of its decisions.

In a user study evaluating saliency maps in image classification, Alqaraawi et al., 2020 found no statistical significance in the presence of classification scores shown with classified samples when predicting model behaviour. However, a study by Verame, Costanza, and Ramchurn, 2016 demonstrated that the presence of scores can encourage the usage of autonomous systems and provide better guidance to users when interacting with these systems. This finding is aligned with the IML cycle, which relies on user feedback to iteratively improve model performance and foster a deeper understanding of model behaviour. By exposing classification scores, the system can enable users to identify which samples to prioritise during the re-labelling process. However, similarly to explanations, system designers should carefully calibrate the level of transparency and detail to optimise user interactions without overwhelming them, particularly non-experts who may struggle to process too much information at once.

4.6.4 Machine Learning and Human Interpretation

4.6.4.1 Balancing Model Performance and Data Reflection

Throughout my experience using TACA, there were numerous moments where I encountered an unexpected conflict between improving the performance of the model and reflecting on the qualitative data. A key assumption made during the design stage of the tool was that, by engaging in the iterative cycle of IML by re-labelling misclassifications and re-training the model, the user would eventually obtain better suggestions from the model and therefore gain more valuable insights from the data.

The participants in the user study did not engage in the cycle sufficiently to assess the validity of the assumption. Conversely, my sustained use of TACA in-

volved several iterations and, in spite of initial biases on model performance, I was convinced by the end of the study that the model had improved, and in fact the average F1 score did actually increase by 0.12. As a result, the suggestions of the model included sentences that I had originally missed when coding the interview transcript. These were definitely valuable in terms of extending the analysis, as I could have used the additional sentences to reinforce additional points or themes that I had originally uncovered. However, perhaps more interesting insights came from reviewing the data I had manually coded. Here, the insights were more about the structural changes to the thematic analysis, suggesting ways I could re-arrange codes and themes to analyse the data from different perspectives.

4.6.4.2 Ground Truth and Human Expertise

My own behaviour around conflicting classifications of ambiguous data was opposite to the one reported by the non-expert participants in the user study. Specifically, as a user with experience in ML and knowledge about the inner workings of the model, I found myself giving less importance to the false positive and false negative classifications in the Confusion Tables compared to the participants in the user study. The difference in how I valued the suggestions of the model compared to the insights gained from reviewing manually coded data could be due to several factors. These classifications were important in the user study, as they presented the participants with a view that was conflicting to their own and were therefore perceived as new perspectives that had not been previously considered. However, I had first-hand experience of the failings of the model during the development and testing phase, so I automatically tended to attribute these classifications to poor model performance, rather than to mistakes in my QDA coding.

Arguably, more work is needed around the ambiguity in the data sets used to train ML models. Unfortunately, the uncertainty, ambiguity, and bias of ground truth data used to train ML models is rarely questioned, even though numerous external factors can significantly influence the data annotation process (Miceli, Schuessler, and T. Yang, 2020). Indeed, reporting on a user study around medical images, Carmichael et al., 2024 pointed to “the complexity of clinical practice

and interpretation, [and] the known imperfection of reference standards” as factors that can introduce ambiguity in ground truth data even in that domain.

In more general terms, even semantically, the terms “false positive” and “false negative” themselves imply that the model is wrong, reinforcing the perception in ML experts that any deviation from the ground truth is a failure of the model rather than calling into question the ground truth (which might not be well-defined). This reflects a broader epistemological issue in ML, which involves overlooking the complexities involved in how training data is created, a topic that has been explored in discussions around the social construction of data sets and their impact on ML outcomes (Bowker and Star, 1999; Miceli, Schuessler, and T. Yang, 2020; Gebru et al., 2022).

Reflecting on and evaluating the ground truth is important for both non-experts and practitioners. For non-experts, it can calibrate trust in the model by increasing awareness of potential training mistakes, helping them avoid viewing the model as an objective, external source of advice, and instead understanding its outputs as contingent on the quality of the training data. For practitioners, it can facilitate the process of improving model performance by identifying labelling inconsistencies before tuning the model hyperparameters.

4.6.4.3 The Role of Machine Learning in Extending Human Analysis

ML does not produce radical new knowledge. Instead, models learn patterns based on the data they have been trained on. As a result, they can only offer suggestions grounded in existing patterns, rather than generating entirely novel insights to challenge the user’s own assumptions. The real value in using ML in analytical processes such as qualitative data analysis lies therefore in its ability to *extend* the user’s existing perspective to unseen data. Only human interpretation, critical thinking, and contextual knowledge can drive the analysis forward with deeper insights or new understanding. Therefore, although complementary or conflicting predictions from the ML model can stimulate deeper engagement with the data, users should be aware that these outputs are not inherently novel insights but rather reflections of

existing patterns in the training data.

In the absence of a well-defined ground truth, human expertise remains necessary as ML cannot replace human judgement. Therefore, tools should be designed with the assumption that users will still need to make critical decisions. In this study, the most interesting insights into the data were gained from reviewing the entirety of the data set, not just the classifications of the model. Enabling users to question the data used to train the model is critical, especially because the quantification of data can lead to a false perception of an external source of knowledge (in non-experts especially).

4.7 Conclusion

This chapter presented an autoethnography on the Thematic Analysis Coding Assistant, an IML tool designed to support qualitative data analysis. The tool was used to extend the thematic analysis performed on the semi-structured interviews from a previous user study on TACA itself. Autoethnography was a particularly suitable research method, as it addressed the limitations of the user study while enabling sustained and genuine use of TACA to analyse and report complementary findings.

The experiences from the use of TACA revealed several findings that were not originally anticipated. Shaped by the roles of developer, researcher and participant, as well as experience in ML, the interactions with the tool differed significantly from those of the participants in every phase of the IML cycle. The design choices in TACA highlighted the importance of balancing transparency and usability to support both efficiency and accuracy in IML systems. The iterative engagement with the model prompted reflections on potential structural changes to the analysis and encouraged reflexivity. As opposed to the experiences of the participants, this was not the result of viewing the model as an external, objective source of advice, but rather a consequence of reviewing both the suggestions of the model and the data used for training.

More generally, this work presented a first-hand account of how ML can be used in analytical processes that do not involve a well-defined ground truth. In this

context, it seems unrealistic to expect ML to produce new knowledge that can challenge existing assumptions. Instead, the true value of ML in analytical processes lies in its ability to extend pre-established perspectives to unseen data. These considerations highlight the need for better communication and education around the capabilities and limitations of ML-driven tools, particularly in how they complement rather than replace human insight.

The findings on model transparency in this chapter suggest that interpretability techniques could significantly enhance the user's contribution to the iterative learning process of the model in IML, aligning with and expanding the insights from Chapter 3. However, the effectiveness of these techniques should first be evaluated in isolation, and evaluation methods in text classification seem lacking, with no studies focused on task performance. Building on these findings, the next chapter presents an evaluation of interpretability techniques for text classification.

Chapter 5

Evaluating Model-Agnostic Interpretability Techniques for Machine Learning Text Classification: A User Study on Predicting Model Outcome

The findings reported in Chapter 3 support the speculation that non-expert participants were subject to the belief of the ML model as an objective source of advice partly due to a lack of transparency of the model (Section 3.5.2). From the responses in the semi-structured interviews, it was clear that the participants did not consider the probabilistic nature of the model and frequently overestimated its performance. Model transparency also plays an important role in the findings of Chapter 4 (Section 4.6.3.2), which suggest how interpretability techniques could play an important role in optimising the user's contribution to the iterative learning process of the model in IML. This chapter builds on these findings to evaluate interpretability in text classification in isolation.

Interpretability is crucial because it directly impacts trust, reliability, robustness, causality, and usability of ML models (Doshi-Velez and B. Kim, 2017), particularly in high-stakes areas such as healthcare, finance, and criminal justice (Samek,

Wiegand, and Müller, 2017). When models are interpretable, stakeholders, including practitioners, regulators, and end users can understand how decisions are made, which is essential to validate the correctness of the model and ensure that it operates fairly and ethically (Langer et al., 2021).

Evaluating these techniques with user studies is critical, as the final goal of interpretability is to improve human understanding, trust, and decision making around ML models. However, the AI research community often prioritises performance metrics as the primary method of evaluating these techniques, potentially overlooking human-centred aspects that are critical to ensuring that explanations are genuinely effective and meaningful in practice. In a recent systematic review on evaluating XAI, Nauta et al., 2023 reported that, among 312 papers published in the past 7 years at major AI and ML conferences, only 22% evaluated with human subjects in a user study.

Within HCI, numerous user studies have been reported on interpretability techniques, measuring self-reported satisfaction, trust, or understanding of model decisions (Lage et al., 2019; Nourani et al., 2019; Papenmeier, Englebienne, and Seifert, 2019; Poursabzi-Sangdeh, Daniel G. Goldstein, et al., 2021b). However, it appears that interpretability techniques in text classifications have not been evaluated through task performance. This chapter aims to address a methodological research gap by conducting a user study in which participants are specifically asked to predict, or anticipate, the behaviour of a text classifier. The study therefore extends the “forward simulation” method (Belle and Papantonis, 2021), which has been applied to different application domains, such as image recognition (Alqaraawi et al., 2020), recommender systems (Scarpato et al., 2024), and decision-support systems (Buçinca, Lin, et al., 2020), to text classification.

The chapter answers the following question:

RQ4: How do interpretability techniques affect users’ ability to predict ML model behaviour?

In particular, the chapter evaluates the two most widely used interpretability techniques. LIME, short for local interpretable model-agnostic explanations, cre-

ates a local surrogate model that closely mimics the prediction of the underlying opaque-box model for a single instance. Unlike the opaque-box model, these local surrogate models are transparent and interpretable. SHAP, short for Shapley additive explanations, explains the prediction of an instance by calculating the contribution of each feature to that prediction using concepts derived from coalitional game theory (Aechtner et al., 2022). The chapter also introduces and evaluates a novel interpretability technique to address existing limitations that uses LLM-generated summaries of the word importance weights generated by LIME.

The study design in this chapter is based on previous work evaluating LRP saliency maps used in Convolutional Neural Network image classification (Alqaraawi et al., 2020), and the RQ is thus deconstructed in a similar manner:

- RQ4.1: Do SHAP and LIME generated word importance heat maps assist participants in predicting the outcome of a text classifier?
- RQ4.2: Are LLM-generated summaries of LIME word importance weights an effective interpretability technique?
- RQ4.3: What are the effects of interpretability techniques on the confidence of predictions of the model outcome?
- RQ4.4: How do different interpretability techniques influence users' attention toward specific features, and what effect does this have on their ability to understand overall model behaviour?

The section below describes the methodology of the user study.

5.1 Method

The between-group study design proposed by Alqaraawi et al., 2020 was adopted to evaluate whether interpretability techniques, specifically LIME and SHAP word importance heat maps and LLM-generated summaries of LIME weights, can help users predict the outcome of a multi-label text classifier. “Multi-label” classification involves predicting zero or more mutually non-exclusive class labels for a given

sample (Tsoumakas and Katakis, 2007). This task was chosen because it allows for the implementation of a model that is more complex than a single-label classifier, while also replicating a more realistic setting where more than one label can be applied to a single text.

5.1.1 Materials

5.1.1.1 Data Set

A collection of 25 restaurant reviews published in the newspaper *The Guardian*¹ between 2022 and 2023 was used as the data set for the study. 25 reviews was the minimum length of the total text (30,000 words) on which the model performed acceptably according to initial tests.

Designing user studies around AI systems is recognised as a challenging task, as it is crucial to ensure positive participant experiences while maintaining ecological validity (Kittley-Davies et al., 2019). Restaurant reviews were relatively engaging, did not require specialised knowledge to understand, did not include insensitive content for participants, and were diverse but shared enough common topics. Moreover, to resemble real-world ML classification scenarios, some level of subjectivity and ambiguity was deliberately introduced in the study by labelling the text using inductive qualitative analysis. This approach reflects the inherent complexity of many real-world data sets, where data does not always fit neatly into well-defined categories and ambiguity between classes (including annotator bias) is common (Geva, Goldberg, and Berant, 2019; Miceli, Schuessler, and T. Yang, 2020).

5.1.1.2 Manual Labelling

The text of the reviews was manually labelled following an inductive qualitative analysis by the researchers in 4 categories or classes: “food and drinks”, “people”, “place”, and “opinions”. The labelling was refined iteratively by training the classifier on the data set and correcting manual mistakes that would result in false positive or false negative classifications, a process also typical of IML (Amershi, Cakmak, et al., 2014).

¹<https://www.theguardian.com/food/restaurants+tone/reviews>

It is noted that subjective and arbitrary choices were made during the analysis, for example, the topics “price” and “menu” belonging to the class label “place”. For clarity, the participants were provided the table of topics manually labelled under each class to the participants during the study (see Table 5.1 for the topic tables for the classes “opinions” and “place”).

Table 5.1: Topics manually labelled belonging to class labels “opinions” and “place”.

Opinions	Place
awards	appearance
personal thoughts	background
positive	decor
negative	layout
reviews	location
	menu
	price

5.1.2 Model Architecture and Training

After manually labelling the data set, all the stop words defined in the NLP library NLTK² were excluded from the labelled samples. Global Vectors for Word Representation (GloVe)³, a learning algorithm to obtain vector representations for words pre-trained on generic Twitter data, were used to generate 50-dimensional word embeddings for each word, before calculating the arithmetic mean of the word vectors for each sample according to the sentence2vec approach⁴.

The study used XGBoost⁵ as the model architecture, an optimised distributed Gradient Boosting library for Python. Gradient Boosting models are considered complex opaque-box models due to their ensemble nature, and are still widely used today for classification and regression tasks (Rudin, 2019; Delgado-Panadero et al., 2022). Additionally, these models are known to work well with smaller data sets (Duan et al., 2020), as opposed to Neural Networks, such as the one used in the study on image classification by Alqaraawi et al., 2020. A Deep Neural Network was also tested by fine-tuning the pre-trained Transformer BERT (Devlin et

²<https://www.nltk.org/>

³<https://nlp.stanford.edu/projects/glove>

⁴<https://github.com/stanleyfok/sentence2vec>

⁵<https://xgboost.readthedocs.io/>

al., 2019), on the same task. However, the state-of-the-art performance of this architecture significantly reduced the occurrence of false positive and false negative predictions, which would not have provided a sufficient number of these cases to sample from to reproduce the study design accurately and gain insights from examples where the model would fail.

The data set was randomly split into training and test sets based on reviews to avoid splitting the same review into different sets. 20 reviews were used in the training set and 5 reviews in the test set. This resulted in a total of 1338 text samples in the training set and 257 samples in the test set. Group K-Fold Cross-Validation was then applied to the training set to identify the best performing hyperparameters for the XGBoost model. The model was then trained on the samples in the training set before being evaluated on the samples in the test set, achieving an average F1 score between the 4 classes of 0.92 on the training set and 0.84 on the test set.

5.1.2.1 Word Importance Heat Maps

The focus was placed on LIME and SHAP, as they are the two most widely used XAI techniques today according to current literature across different domains (Aechtner et al., 2022; Cesarini et al., 2024; Salih et al., 2024). Furthermore, model-agnostic interpretability techniques allow comparisons across different model architectures (Adadi and Berrada, 2018), extending the findings of the study beyond Gradient Boosting classifiers.

The Python library ELI5⁶ was used to generate LIME weights and SHAP⁷ was used to generate SHAP weights. Weights are associated with each word in each data sample and range from negative to positive values, depending on whether the word has a positive or negative impact towards the classification of an individual class. The word importance heat maps were generated using seaborn⁸, a statistical data visualisation library. Replicating the colour palette of LRP saliency maps, red was used to indicate words that support the classification, while blue indicated words that go against the classification. Figure 5.1 shows the same true positive example

⁶<https://eli5.readthedocs.io/>

⁷<https://shap.readthedocs.io/>

⁸<https://seaborn.pydata.org/>

for the class “opinions” with heat maps generated from LIME and SHAP weights.

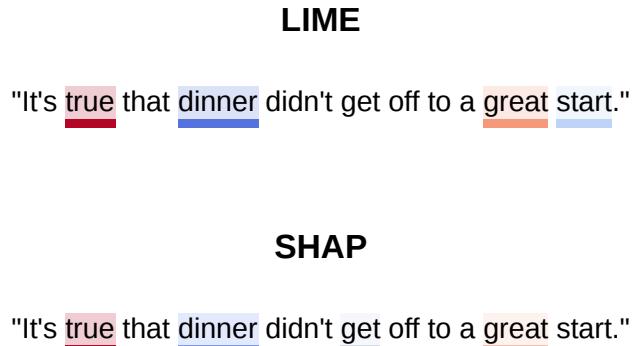


Figure 5.1: Word importance heat maps generated from LIME and SHAP weights according to the class “opinions”.

5.1.2.2 LLM-Generated Summaries

The limitations of local, example-based explanations have been widely recognised and discussed in the literature. These techniques generate explanations that are relatively easy to understand, but only capture the behaviour of the model within a specific local region of the input space (Ribeiro, Singh, and Guestrin, 2018). Abstracting local insights to a global understanding can be challenging as it is often unclear whether the explanations produced by these techniques are applicable beyond the specific instance for which they were generated (Alqaraawi et al., 2020; Chromik et al., 2021). To address these limitations, a novel interpretability technique was designed to capture the behaviour of the model on the entire input space.

Recent advances in LLMs have significantly improved their ability to summarise text efficiently (Luo, Xie, and Ananiadou, 2023), as Automatic Text Summarisation (ATS) has drawn considerable interest in both academic and industrial circles (Jin et al., 2024). Recent work from X. Yang et al., 2023 revealed that the performance of ChatGPT in summarisation tasks is remarkable, as indicated by Rouge scores, showing that it can often rival traditional fine-tuning methods.

With this in mind, words from the entire training data set were summarised according to their LIME importance weights using ChatGPT 4o. LIME weights were selected over SHAP weights because an initial pilot study revealed that these

performed slightly better in predicting the model outcome. The trained XGBoost model was run on the training set, and a list of all words and LIME weights from all the predicted training samples was extracted, using the average weight for words appearing more than once. The lists for the class label “opinions” each contained 4351 pairs of terms and weights (ranging from -1.83 to 1.82 in “opinions” and -2.58 to 1.86 in “place”). The table was then uploaded to ChatGPT (see Appendix C.5.3 for the prompt used).

The focus was particularly on summarising words with a strong positive weight, neutral weight, and negative weight, to reflect the word importance heat maps in LIME and SHAP. 250 words was about the average length of the examples in the other three conditions. During a testing phase, it was observed that longer summaries were only more verbose and did not mention additional topics or words. See Appendix C.5.1 and C.5.2 for the summaries generated for the classes “opinions” and “place” used in the study.

The code used to train the text classifier, sample examples and tasks, and generate heat maps is released as open source⁹.

5.1.3 Tasks

The participants were shown 12 classification **examples** and asked to identify words or topics to which the system is sensitive, as well as those not considered by the system. Participants were also asked whether *they* would assign the class label to the task sentence immediately before predicting the outcome of the model. This question was included because in pilot studies following the think-aloud protocol, it was noticed that participants would often make predictions based on whether they would assign the label themselves.

As the main task of the user study, the participants were then asked to predict whether the model will assign the class label to a given **task** data sample. Participants were also asked to rate their confidence in the prediction using a 4-point Likert scale. This was repeated 14 times, each time with a different set of examples and task: 7 times for the label “opinions” and 7 times for the label “place”. These

⁹<https://github.com/fmilana/explanations>

classes were chosen because the model performed best and worst on them, respectively (0.84 and 0.78 F1 scores on the test set), and had enough false negative and false positive classifications to sample examples from. To avoid order bias, half of the participants were shown the “opinions” samples first, while the other half started with “place”. At the top of each page, before the examples and questions, participants were also shown the topics that belonged to the class label (see Table 5.1). See Appendix C.6 for an example of a single page from the study with LIME weight importance heat maps.

At the end of the study, participants exposed to LIME and SHAP were also asked whether the example categories (e.g. false positive) or the word importance heat maps, or both, were more useful when predicting the model outcome.

5.1.3.1 Selection of Examples

The examples were selected from the training data set. This decision was made because these samples represent the data that the model has already learned patterns from, which would allow the participants to observe cases that align with the understanding of the model. The examples were sampled based on their cosine vector distance to the corresponding task sample, ensuring that participants were shown examples conceptually similar to the sample on which they were asked to predict model behaviour. Displaying the outcomes of the classifier has been shown to be important for the effectiveness of explanation methods (Lai and Tan, 2019). Therefore, the examples were categorised as true positive, false negative and false positive, replicating the study design on image classification (Alqaraawi et al., 2020) according to the following distribution:

- **6 true positive examples:** Samples that were manually labelled as class X and that the model correctly classified as X
- **3 false negative examples:** Samples that were manually labelled as class X and that the model incorrectly did *not* classify as X
- **3 false positive examples:** Samples that were *not* manually labelled as class X and that the model incorrectly classified as X

5.1.3.2 Selection of Tasks

The pilot studies revealed that the time spent on each task prediction was the same as that of Alqaraawi et al., 2020. Therefore, the same number of tasks was sampled: 7 for each of the 2 classes. In the same fashion, 3 true positive, 2 false negative, and 2 false positive samples were selected for each class to include cases where the model outcome was both correct and incorrect. All tasks were selected from the test set. Drawing samples from the test set allowed for an evaluation of the participants' ability to predict model behaviour on new, unseen data, reflecting the real-world application of ML models. The samples were based on their classification score, or probability, according to the model. Since the prediction task should not have been too easy or too difficult, the samples were restricted to have a midpoint classification score of around 0.75 for true positive and false positive samples and 0.25 for false negative samples (the classification thresholds were set at 0.5 for all classes, the default value used in binary classification tasks). See Appendix C.7 for a list of all the tasks selected.

5.1.4 Conditions

The study measured **presence and type of interpretability technique** as the main independent variable. This factor has 4 levels:

- **No explanations** (baseline level): examples shown without word importance heat maps
- **LIME**: LIME word importance heat maps shown on the examples (*not* on the tasks)
- **SHAP**: SHAP word importance heat maps shown on the examples (*not* on the tasks)
- **Summaries**: LLM-generated summaries shown of LIME weights of all samples in the training data set *instead* of the examples

A secondary independent variable was involved in the analysis to provide additional insights on the interaction effects of interpretability techniques: task category,

i.e. the category of the sample on which participants predicted the model outcome. This factor has 3 levels (the participants were unaware of the task category):

- True positive task
- False positive task
- False negative task

Alqaraawi et al., 2020 found no statistical significance in the presence of classification scores, so that factor was excluded from this study.

5.1.5 Participants

A total of 128 participants (32 for each condition level) was recruited from *Prolific*¹⁰, an online crowd-sourcing recruitment platform. The recruitment criteria were: minimum age of 18, fluency in English and a degree in a technical subject (e.g., mathematics or engineering) to ensure that participants were more familiar with technical terms such as “false positive”. Participants were also required to have an approval rate of at least 95% on Prolific to ensure the quality of the responses. 56 of the 128 recruited participants stated that they had already learned about ML (44%), and 31 had already worked with ML (24%).

5.1.6 Procedure

The study was hosted on Qualtrics¹¹. Participants received the study instructions (see Appendix C.3 and C.4) after reading the study information sheet (see Appendix C.1) and giving their informed participation consent (see Appendix C.2). The instructions explained the basic notions of ML: the training and test data sets and the classification task, as well as the classes used to manually label the restaurant reviews. The instructions also provided detailed explanations of the terms “true positive”, “false positive”, “false negative” and “true negative”, and a description of the heat maps or LLM-generated summaries, if present, including an example. After receiving instructions on the user study task, participants were given 2 chances to

¹⁰<https://www.prolific.co/>

¹¹<https://qualtrics.com/>

answer a comprehension check correctly. The check required participants to identify a classification example belonging to the “false negative” category, and whether individual words contributed towards or against the classification, if in the LIME or SHAP condition level. Participants who did not answer the comprehension check correctly twice were asked to return the study submission. Participants were compensated for their time (40 minutes) with £6, in compliance to the local living wage. The study was reviewed and approved by the Ethics Committee of UCLIC (Project ID No: UCLIC/1617/017/Staff Costanza/Nowacka/Yang).

5.1.7 Pilot Studies

The methodology described above was the product of refinement based on the insights gathered from two initial pilot studies, each involving 64 participants (16 participants in each group).

5.1.7.1 Pilot Study 1

When initially designing the study, it was incorrectly assumed that reading text would take longer than going through the images in the study from Alqaraawi et al., 2020. To keep the study length comparable and avoid fatigue effects, the participants in the No-Explanation, LIME and SHAP groups were shown 8 true positive, 4 false positive, and 4 false negative examples, and then asked to predict model behaviour on a batch of 4 tasks. The process was then repeated for the second class, for a total of 8 tasks (compared to 14 in the final study design).

The sampling of tasks was considerably different. The tasks were sampled on the predicted set and not the test set. The tasks were also not calculated on vector distance from the examples. Instead, these were sampled following the distribution:

- 1 task near the minimum classification score
- 1 task near the first quartile of the classification scores
- 1 task near the third quartile of the classification scores
- 1 task near the maximum classification score

Although the intention was to provide varying levels of difficulty for the prediction task, the result was that the majority of tasks were too easy to predict: out of 8 total tasks, 3 were predicted correctly by at least 90% of participants, 6 by at least 80%. In this context, the heat maps could provide little to no value. The results of Pilot Study 1 are reported in Appendix C.8.1.

5.1.7.2 Pilot Study 2

A second pilot study was run to sample tasks that were more difficult to predict. This time, the tasks were sampled from the test set so that categories could be taken into consideration too. Each class now included 2 false positives and 2 false negatives, as these categories were assumed to increase the difficulty of the prediction. The tasks were also sampled closer to the classification threshold, following the distribution:

- 1 false positive task near the bottom quartile of the classification scores
- 1 false positive task near the median between the threshold and the bottom quartile of the classification scores
- 1 false negative task near the median between the threshold and the top quartile of the classification scores
- 1 false negative task near the top quartile of the classification scores

As shown in Appendix C.8.2, the tasks were now too difficult. Due to chance, the results actually showed a statistical difference in prediction accuracy after 32 participants took part in the pilot, but the final results did not reveal any difference on prediction based on conditions.

After this pilot study, the design was changed to replicate the one from Alqaraawi et al., 2020 more closely by: sampling the tasks based on cosine vector distance from the examples, sampling 2 true positives, 1 false positive and 1 false negative for each class, and asking participants to predict each of the 14 tasks separately after being shown a set of 12 examples. The difficulty in task prediction was also attributed to recognisable mistakes in the manual labelling of the training data set. These were corrected following the process described in Section 5.1.1.2.

5.2 Results

5.2.1 Outcome Prediction Accuracy

The Aligned Rank Transform (ART) method, a non-parametric approach to factorial ANOVA (Wobbrock et al., 2011), was used, as Levene’s test indicated that the data did not meet the assumption of homogeneity of variances ($F(11,372) = 2.29, p = 0.01$). This method was applied to reveal the effects of the condition (No-Explanation, LIME, SHAP, and Summaries) and task category (true positive, false negative, and false positive) on the accuracy of the predictions.

The ART ANOVA did not reveal a significant main effect of condition on prediction accuracy ($F(3,124) = 0.87, p = 0.46$). The test revealed a significant main effect of the category ($F(2,248) = 61.81, p < 0.001, \eta_p^2 = 0.25$), indicating a medium to large effect size on accuracy. Additionally, a significant interaction effect between condition and category was found ($F(6,248) = 2.70, p = 0.01, \eta_p^2 = 0.04$), representing a small effect size for the interaction.

Post-hoc pairwise comparisons with Holm adjustment using the ART-C method were conducted to explore the significant main effect of category ($\eta_p^2 = 0.25$). The prediction accuracy on true positive samples was significantly higher than on false negative samples ($mean = 66.9\%, SD = 21.5\%$ vs. $mean = 44.5\%, SD = 27.9\%, p < 0.001$) and on false positive samples ($mean = 32.0\%, SD = 26.2\%, p < 0.001$). The accuracy on false negative samples was also significantly higher than on false positive samples ($p < 0.001$).

The ART-C post-hoc analysis revealed significant interaction effects between the condition and category on prediction accuracy, but these were only where both the condition and the category were different. Refer to Figure 5.2 for the distributions of the total correct predictions.

Participants were asked to predict the model outcome for a total of 14 tasks. A chi-square test was performed to reveal differences in prediction accuracy *based on the individual samples* (see Appendix C.7). The test revealed a significant difference for 2 of the 14 samples. The first is a false negative sample for the class label “place” ($\chi^2 = 20.65, p < 0.001, df = 3$, Cramér’s $V = 0.40$):

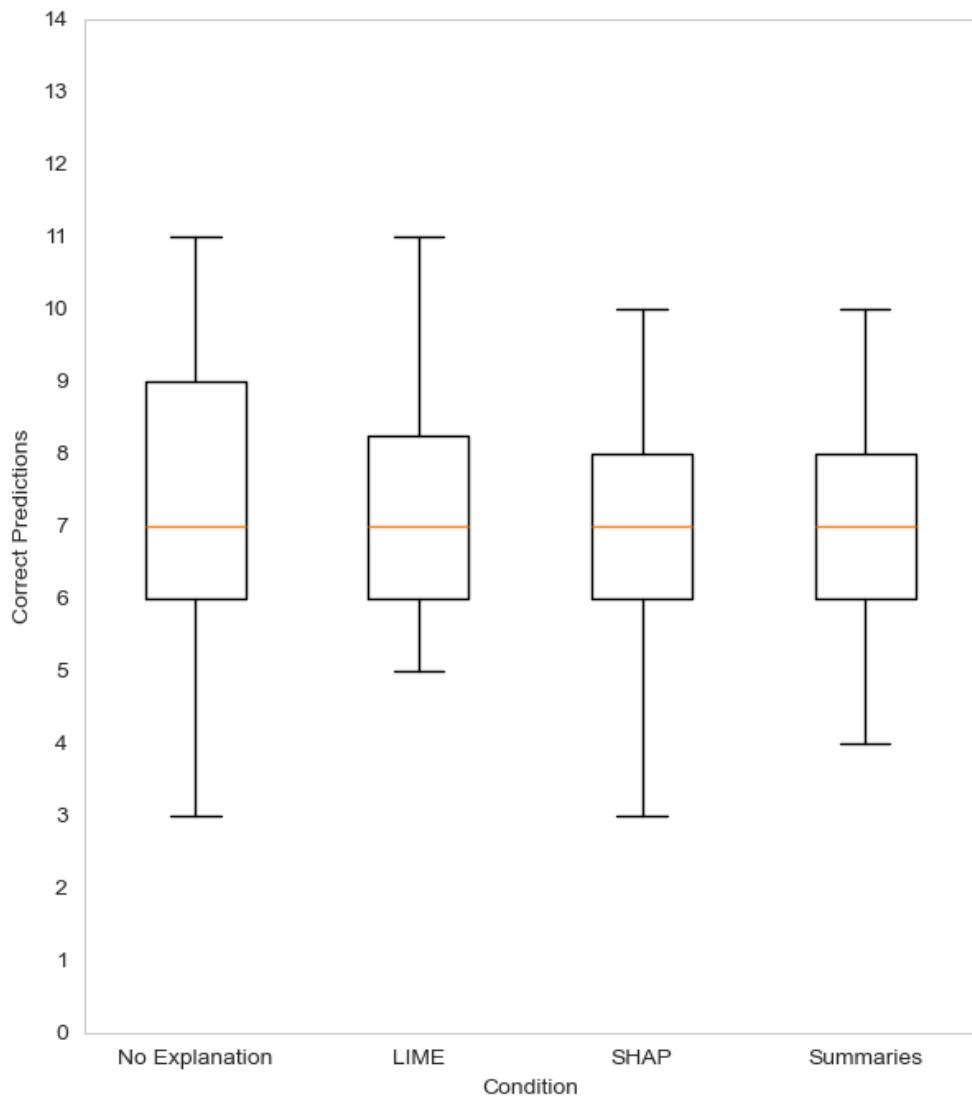


Figure 5.2: Distributions of total correct predictions to all questions by condition.

“The trout dish is £24.”

For this specific task, an Adjusted Standardised Residual post hoc test revealed fewer correct predictions than expected in the No-Explanation condition (*adjusted residual* = -2.16), and more correct predictions than expected in the summaries condition (*adjusted residual* = 2.41).

The test also revealed a significant difference ($\chi^2 = 4.37, p = 0.22, \text{df} = 3$, Cramér’s $V = 0.19$) in prediction accuracy for the following false positive task for the class label “place”:

“Mendes is also the executive chef.”

The post hoc test revealed more correct answers than expected for this task in the No-Explanation condition (*adjusted residual* = 2.3).

5.2.2 Confidence in Predictions

On a scale of 1 to 4, participants tended to be “slightly confident” in their predictions on average (*median* = 3) across all conditions. See Figure 5.3 for the distributions of the average confidence scores.

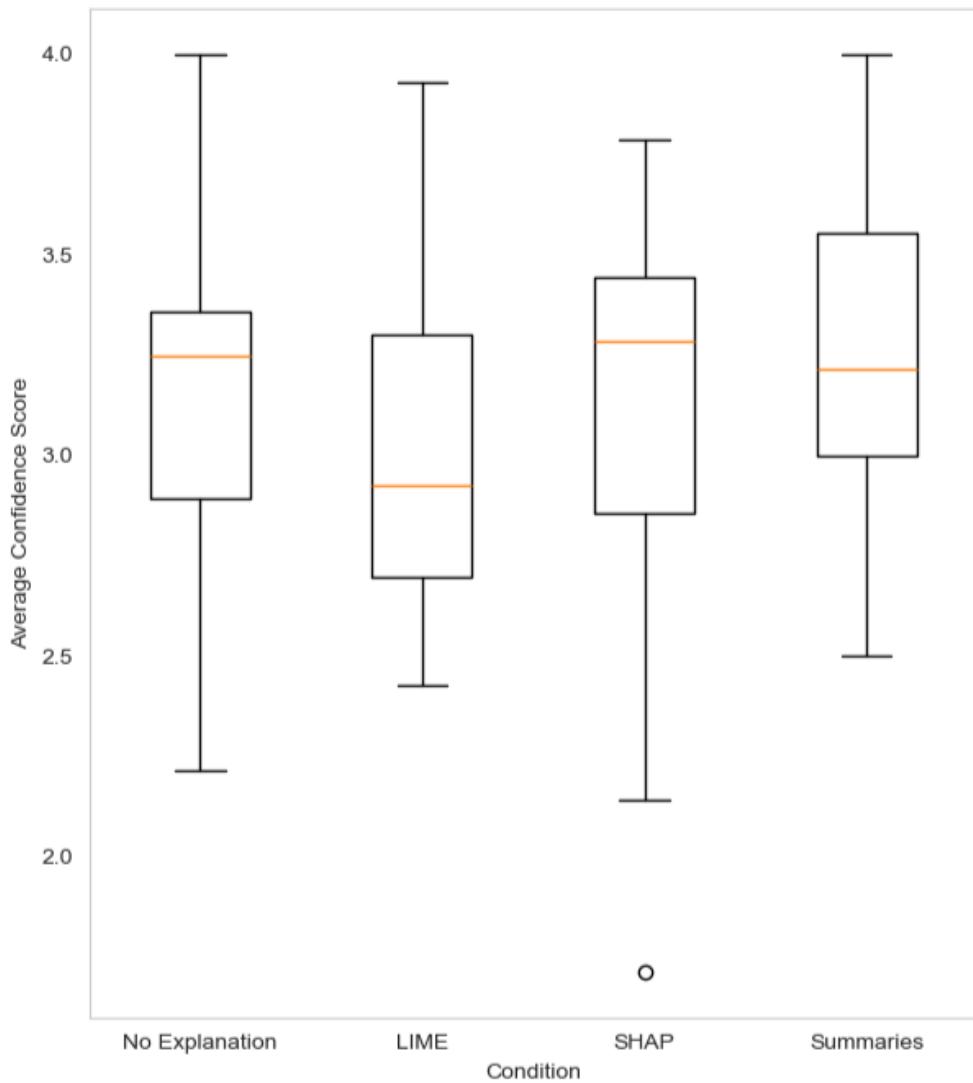


Figure 5.3: Distributions of average confidence scores (1-4) to all questions by condition.

The Shapiro-Wilk tests for normality indicated that all groups significantly deviated from a normal distribution (e.g., No-Explanation: $W = 0.96, p = 0.005$; LIME: $W = 0.95, p = 0.001$). Due to these violations, the ART ANOVA was run to reveal the effects of the condition and task category on the confidence of the participants in their outcome predictions. The test did not reveal a significant main effect of the condition on confidence ($F(3, 124) = 2.11, p = 0.10$). The test revealed a small significant main effect of category ($F(2, 248) = 5.21, p = 0.006, \eta_p^2 = 0.01$), and a small significant interaction effect between condition and category ($F(6, 248) = 2.55, p = 0.02, \eta_p^2 = 0.02$).

Post-hoc pairwise comparisons with Holm adjustment using ART-C revealed that confidence was higher for predictions of false negative samples than true positive samples ($mean = 3.20, SD = 0.53$ vs. $mean = 3.08, SD = 0.52, p = 0.006$). Only one specific comparison—between the confidence in true positive predictions using LIME and false negative predictions using Summaries was statistically significant ($mean = 2.92, SD = 0.43$ vs. $mean = 3.34, SD = 0.50, p = 0.04$). Chi-square tests on individual tasks revealed no statistical significance.

5.2.3 Perceived Usefulness of Example Categories and Word Importance Heat Maps

When asked whether the example categories or the word importance heat maps (or both) were most useful, 5 participants in LIME selected only the categories, 13 participants selected only the heat maps, and 14 selected both. 7 participants in SHAP selected only categories, 11 participants selected only the heat maps, and 14 selected both. A chi-square test did not reveal significant differences between the two condition levels.

5.2.4 Mentioned Features

Before making a prediction of the outcome of the model, participants were asked to identify 1-3 features that the model was sensitive to and those that it ignored, if they found any (the answers were free text). These questions were made before every prediction in the No-Explanation, LIME, and SHAP conditions (14 times),

and before the first prediction made for each class label in the Summaries condition (2 times), as only one summary per class label was included.

5.2.4.1 Mixed-Methods Analysis

A qualitative analysis was conducted on all the responses that participants gave to questions related to features. This involved several passes of inductive coding leading to the three code groups: **terms**, **topics**, and **language**. Each separate concept submitted to the free text questions was assigned to one of these groups.

In **terms**, all words included were those taken directly from the examples or summaries. These could be words highlighted by the LIME and SHAP word importance heat maps or mentioned as examples in the LLM-generated summaries (e.g. “good”, “London”, or “cheerful”). In **topics**, words were included that referred to concepts abstracted by participants from the examples in the No-Explanation, LIME, and SHAP conditions, or were possibly mentioned by the summaries in the Summaries condition (e.g. “personal thoughts”, “location”, or “appearance”). The words in **language** were those that referred to the style, tone, or specific linguistic features of the text (e.g. “specific/descriptive”, “adjectives”, “names”). See Figure 5.4 for the average frequency of terms, topics, and language feature types per question across conditions.

Kruskal-Wallis tests¹² revealed statistical differences between conditions on the frequency of **terms** ($H(3) = 29.94, p < 0.001, \eta_p^2 = 0.22$), **topics** ($H(3) = 25.63, p < 0.001, \eta_p^2 = 0.22$), **language** ($H(3) = 8.55, p = 0.04, \eta_p^2 = 0.11$), and the total number of features ($H(3) = 26.97, p < 0.001, \eta_p^2 = 0.19$). The following frequencies were calculated per question.

Regardless of the feature type, participants in the No-Explanation condition mentioned significantly fewer features ($mean = 1.86, SD = 0.32$) than participants in LIME ($mean = 2.43, SD = 0.48$) ($p = 0.001$), SHAP ($mean = 2.54, SD = 0.46$) ($p < 0.001$), and Summaries ($mean = 3.64, SD = 3.35$) ($p < 0.001$). Participants in Summaries mentioned more features than participants in LIME ($p = 0.004$) and in

¹²Kruskall-Wallis was run here instead of a factorial test like ART ANOVA as the task category was not of interest as an additional factor. The participants were asked to mention features based on the examples/summaries.

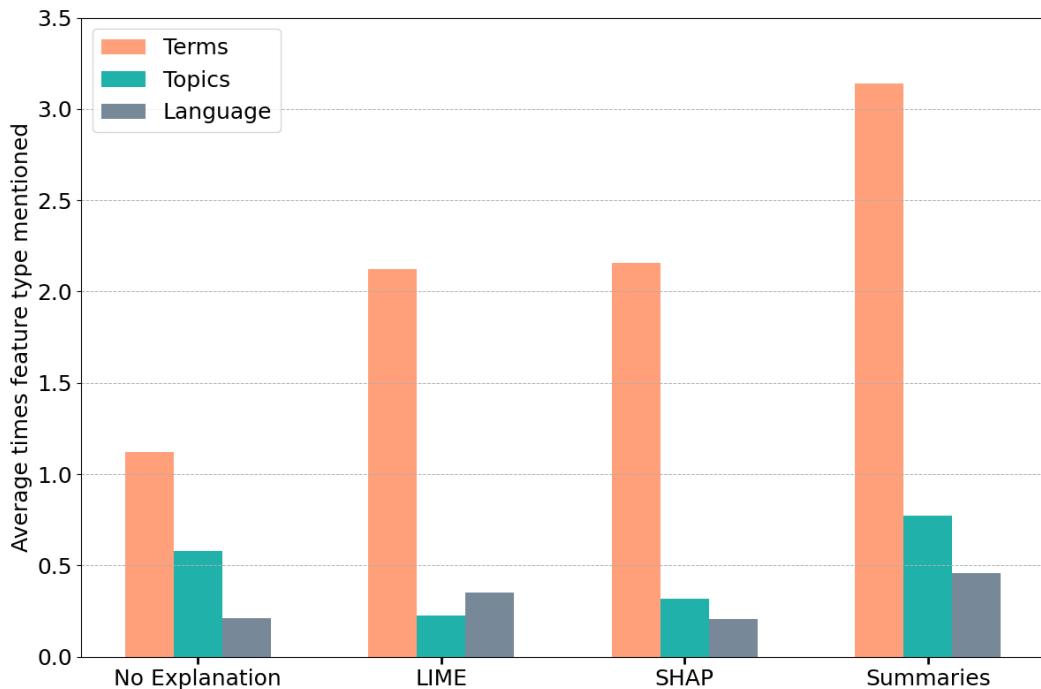


Figure 5.4: The average frequency of terms, topics and language feature types by condition.

SHAP ($p = 0.007$).

More specifically, participants in Summaries mentioned more **terms** than those in No-Explanation ($mean = 3.14, SD = 4.2$ compared to $mean = 1.12, SD = 0.69, p < 0.001$), LIME ($mean = 2.12, SD = 0.87, p = 0.04$) and SHAP ($mean = 2.16, SD = 0.78, p = 0.04$). Participants in LIME and SHAP mentioned more **terms** than participants in No-Explanation ($p < 0.001$ in both cases).

Participants in Summaries mentioned more **topics** than participants in LIME ($mean = 0.78, SD = 0.45$ compared to $mean = 0.22, SD = 0.05, p < 0.001$) and SHAP ($mean = 0.32, SD = 0.19, p < 0.001$) but *not* No-Explanation ($mean = 0.58, SD = 0.26, p = 0.18$). Participants in No-Explanation mentioned more **topics** than participants in LIME ($p = 0.003$) and those in SHAP ($p = 0.02$).

In terms of **language**, participants in Summaries mentioned more of these features than participants in No-Explanation ($mean = 0.46, SD = 0.07$ compared to $mean = 0.21, SD = 0.04, p = 0.01$) and SHAP ($mean = 0.21, SD = 0.03, p = 0.02$) but *not* LIME ($mean = 0.35, SD = 0.07, p = 0.25$).

See Figure 5.5 for the most frequently mentioned salient features for both class

labels across conditions.



Figure 5.5: Normalised frequencies of salient features mentioned by participants for samples about “opinions” (top 2) and “place” (bottom 2). Left: specific terms, right: topics and language.

5.3 Discussion

After a combination of quantitative and qualitative analysis, the results of this study reveal several key aspects of model-agnostic interpretability techniques applied to text classification. The following subsections discuss the lack of effects of these techniques on model outcome prediction and confidence, how they still influence users to notice more specific features of the text, and the broader implications of the

findings.

5.3.1 No Effects of Techniques on Model Outcome Prediction

The literature in XAI has often evaluated interpretability techniques by measuring one or a combination of performance metrics (Atanasova et al., 2020; Cesarini et al., 2024), user understanding (Cesarini et al., 2024), self-reported trust (Nourani et al., 2019; Papenmeier, Englebienne, and Seifert, 2019; Cesarini et al., 2024), human-grounded benchmarks (Atanasova et al., 2020; Mohseni, Block, and Ragan, 2020). Instead, the design of this study followed the reasoning of Muramatsu and Pratt, 2001 to measure outcome prediction and adopted the evaluation method used in related work (B. Kim, Khanna, and Koyejo, 2016; Alqaraawi et al., 2020; Buçinca, Lin, et al., 2020; Waa et al., 2021) on different task domains to extend the findings to text classification. The findings reveal that LIME and SHAP word importance heat maps and LLM-generated summaries of LIME weights from the training data set samples did not help participants anticipate model behaviour. Still, the majority of participants in this study considered the LIME and SHAP heat maps useful to predict the outcome of the model (84.4% and 78.1%, respectively).

5.3.1.1 Prediction Accuracy

On average, LIME and SHAP heat maps did not have a measurable effect on the accuracy of the predictions in this study. These findings are in line with previous research comparing the effects of rule-based and example-based explanations in decision support systems (Waa et al., 2021), which concluded that both types of explanations are insufficient alone to significantly improve task performance. They are also in line with work from Buçinca, Lin, et al., 2020, which found that while explanations increased user trust and preference in another decision-support system, they did not improve actual decision-making performance. However, this result is in contrast to that of the same evaluation method applied to image classification (Alqaraawi et al., 2020), where LRP saliency maps produced a significant positive effect on task performance, despite the success rates remaining relatively low (60.7% with the saliency maps compared to 55.1% without the saliency maps).

It could be argued that the difference in results between image and text classification can be simply explained by the quality of the explanations. Model-specific methods such as LRP have been found to perform better than model-agnostic methods (Atanasova et al., 2020). However, the design of this study prioritised generalisability across all types of models; gradient boosting was chosen due to the small size of the training data set, even though Transformers are currently more commonly used for text classification. Additionally, LIME and SHAP remain the most popular interpretability techniques and are still widely used in this domain (Aechtner et al., 2022; Cesarin et al., 2024; Salih et al., 2024).

Arguably a more plausible reason for the difference between the effects of explanations in image and text classification is the concept of “unclear coverage” mentioned by Ribeiro, Singh, and Guestrin, 2018, which refers to the ambiguity and limitations associated with local explanations. More specifically, it is generally unclear whether the explanations generated by these techniques apply beyond the specific instance for which they were generated, possibly misleading users in predicting model behaviour in new situations. The number of human-identifiable features in images is considerably lower than in text, which could help users abstract higher-level concepts and patterns from the examples, therefore increasing “coverage”. In fact, participants in the image classification study mentioned features such as “eyes”, “ears” and “legs” which are significantly more likely to reappear in task image samples than individual terms such as “good”, “explain” and “menu” in task text samples.

The “explanations by example” from LIME and SHAP may not be sufficient to improve task performance in complex domains such as text classification. There are several implications for the design of interpretability techniques. First, that example-based local explanations are not enough to capture the nuances of complex data structures found in text, and could likely benefit from a combination with rule-based global explanations, as discussed in previous work on other data types (B. Kim, Khanna, and Koyejo, 2016; Scott M. Lundberg and S.-I. Lee, 2017; Waa et al., 2021). The contrast between image and text classification also suggests that

interpretability techniques need to be tailored to specific tasks depending on the number of identifiable features. Techniques that work well on fewer and more tangible features like in images may not translate effectively to tasks with more abstract features, such as text. This observation aligns with work by Poursabzi-Sangdeh, Daniel G Goldstein, et al., 2021a, which shows that model transparency is highly dependent on the number of features involved.

The findings also confirm the importance of evaluating techniques on task performance rather than relying only on subjective measures such as perceived usefulness, as most participants reported finding the heat maps useful despite not actually making more accurate predictions. This indicates that some metrics used for evaluation in the literature may not completely measure the real effectiveness of these techniques when applied in real-world decision-making scenarios.

In an attempt to address the potential issue of “unclear coverage” of local explanations, the study also included LLM-generated summaries of LIME word importance weights that could provide a global approximation of the behaviour of the model. However, the summaries also failed to significantly increase the accuracy of the predictions. The lack of effects of LLM-generated summaries on prediction accuracy suggests that, while they offer a global perspective on model behaviour, they may also oversimplify the complexity required for accurate decision-making. By abstracting information, summarisation is likely to omit essential details, which can mislead users, particularly in text classification where data is rich and variable.

This issue is illustrated by the role of the topic “price” in this study. Text samples related to price were manually labelled in the category “place”. However, these were only a small subset compared to the samples about location or hospitality, so the words related to price with a positive weight in the list extracted from LIME also appeared very infrequently. As a result, the summary of “place” contained no mention of prices (see Appendix C.5.2). An effect of this can be seen in the prediction accuracy of the task (“The trout dish is £24”), on which more participants in the Summaries condition predicted a negative model outcome. The increased accuracy of participants in Summaries for this individual task is likely fortuitous. The model

should have recognised the price but did not (false negative). However, participants probably predicted a negative outcome not because they understood the behaviour of the model, but because they had no information regarding prices.

The insight into the role of “price” in prediction accuracy emerged largely because the study deliberately introduced ambiguity and subjectivity in the data set during the manual labelling process. Had the relationship between the topic and the class been obvious, this nuanced effect on prediction accuracy might not have been observed. Such complexity mirrors real-world scenarios where the ground truth is not always clear-cut and can lead to unexpected effects on model behaviour and user interpretations. User studies on ML should account for these complexities to provide more realistic assessments of user behaviour and more actionable results.

Overall, the findings suggest that summarisation of word importance weights should be carefully designed to avoid presenting the same issue of “unclear coverage” in local explanations. Currently, summaries of words based on their importance weights seem to be as ineffective in improving task performance as conventional rule-based explanations (Waa et al., 2021). Future work could evaluate hybrid approaches that combine LLM-generated summaries with either local, example-based explanations, or interactive elements enabling users to explore specific features in more depth.

5.3.1.2 Prediction Confidence

The results indicate that none of the explanation types have an effect on the confidence in the prediction of the model outcome. Participants were “slightly confident” in all conditions. This finding is in contrast with the results of previous studies that explanations give users a false sense of confidence by overestimating their understanding of model behaviour (Rozenblit and Keil, 2002; Bussone, Stumpf, and O’Sullivan, 2015; Schaffer et al., 2019) and the concept of “illusion of explanatory depth” (IOED), the phenomenon where users feel more confident in their understanding due to the presence of explanations (Rozenblit and Keil, 2002; Collaris, Vink, and Wijk, 2018; S. T. Mueller et al., 2019; Sokol and Flach, 2020; Chromik et al., 2021).

The discrepancy could be due to several factors. One possibility is that, unlike the studies mentioned above, the participants in this study had a technical background and almost half of them had already learned about ML, which could have contributed to a more cautious and critical evaluation of the explanations provided. This would indicate that the potential for explanations to induce a false sense of confidence may be mitigated when users have a stronger foundation in the underlying technology.

Interestingly, participants were more confident in their predictions of false negative samples than true positive samples. However, the prediction accuracy on false negative samples was significantly lower than on true positive samples. This suggests that certain features or patterns in the false negative samples were misleading. For example, LIME and SHAP might have highlighted certain words in red that participants believed were strongly indicative of the given class label.

This misalignment between confidence and task performance reveals the significance of designing techniques that not only help users understand the decisions of the model but also guide them in evaluating their own understanding more accurately. If explanations inadvertently reinforce incorrect assumptions or emphasise features that are not truly indicative of the decision-making process, they can lead to overconfidence in incorrect predictions. Additionally, the findings present an opportunity to further investigate the effects of explanations, whether local or global, on user confidence in relation to the technical knowledge of the users.

5.3.2 Effects of Techniques on Feature Attention

On average, participants who were exposed to any of the three interpretability techniques evaluated mentioned significantly more salient and ignored features. This finding is in line with similar work on image classification (Alqaraawi et al., 2020). However, since these participants did not experience higher prediction accuracy, it also confirms that transferring knowledge about potential features to new samples, where they might appear in a different context, is challenging (Chromik et al., 2021).

Participants in the LIME and SHAP condition levels were particularly drawn to

individual terms, more so than those in No-Explanation. The results clearly indicate that participants paid attention to the terms highlighted by the heat maps in red and blue. In contrast, participants who did not see any explanations were more prone to mention topics as salient or ignored features. Both of these findings are consistent with observations reported in studies involving image classification (Alqaraawi et al., 2020).

The terms most frequently mentioned by the participants exposed to heat maps include words like “good”, “feel”, “explain” and “profound” for the class label “opinions”, and “menu”, “restaurant”, “place” and “town” for the class label “place” (see Figure 5.5). Notably, none of these words are present in the respective task examples (see Appendix C.7). In the context of interpretability, Chromik et al., 2021 argue that people overestimate their ability to accurately recall observations, leading to discrepancies between stored mental images and the original facts. Additionally, most complex systems are hierarchical, and if people can name and describe individual parts on the first level of the hierarchy, they often assume to understand how the overall system works. The participants in the explanation condition levels of this study did not report higher levels of confidence, but this assumption could have driven them to overlook broader patterns in the data.

The fact that word importance heat maps draw user attention away from general attributes of samples, like topics or language, is significant because these attributes might help users extract more transferable knowledge of model behaviour. Arguably, by focusing too narrowly on specific highlighted terms, users could have missed general patterns or concepts that were likely to reappear in the tasks (at least more frequently than the individual terms). An opportunity for future work is to investigate further how participants reason to predict tasks based on the examples, for example through a think-aloud protocol or pair studies that could provide meaningful qualitative data to test this hypothesis.

Unsurprisingly, summaries elicited participants to mention more topics than those in the LIME and SHAP conditions. These were mostly limited to the ones mentioned in the summaries themselves, as participants hardly ever generalised to

wider or different patterns. Selective attention to the topics most frequently mentioned in the summaries could have led these participants to miss other potentially important features present in the tasks. For example, the task for the class label “opinions” with the lowest prediction accuracy from participants in Summaries (0.09) includes words with a positive connotation: “pleasingly”, “pretty” and “perfect” (see Appendix C.7). These terms are in line with the salient topics mentioned in the summary, “positivity and emotional well-being”, and in fact participants mentioned “positive” and “thoughts” as salient features and the majority (0.91) predicted a positive model outcome. However, the task was actually a false negative, most likely because the vast majority of words, such as “anchovies”, “tomatoes”, and “eggs” describe specific details related to food, and the few positive words failed to push the average embedding toward the space where the entire sentence would be classified as “opinions”.

These findings reinforce the need to design interpretability tools and techniques that not only indicate important features, but also encourage users to consider broader context and patterns. For example, B. Kim, Wattenberg, et al., 2018 introduced an interactive technique to enable users to test and explore the influence of various high-level concepts on the classification result. Building on this work, future research could be focused on tools that allow users to combine local and global explanations to develop a holistic understanding of the decision-making process of the model.

5.4 Conclusion

This chapter reported on a between-group user study aimed to evaluate the two most widely used model-agnostic interpretability techniques, LIME and SHAP, as well as a novel summarisation technique based on LIME-weighted words, in text classification on a data set consisting of labelled newspaper restaurant reviews. This work replicated a similar study on LRP saliency maps on CNN image classification, where the technique was evaluated based on how accurately participants were able to predict model outcome.

The results revealed that, unlike LRP saliency maps in image classifications, the interpretability techniques evaluated in text had no substantial effects on the precision or confidence in model outcome prediction. When asked to mention features that the model found salient and ignored, participants with no explanations identified more topics, those with LIME and SHAP heat maps identified more terms, and those with Summaries identified more features overall. However, the features mentioned with explanations were not particularly indicative of model behaviour.

The importance of evaluating techniques on task performance, rather than solely on user feedback, is a crucial takeaway from this study. Although users may perceive certain explanations as helpful, this does not necessarily translate into better outcomes. Future research should continue to explore metrics that can accurately capture the effectiveness of interpretability techniques in real-world applications.

This study contributes to the growing body of literature on interpretability in AI by revealing the limitations and potential pitfalls of current techniques in the context of text classification. Example-based explanations provide valuable insights into model behaviour, but they should be complemented by global, possibly interactive explanations. These could help users grasp not only specific features influencing a single prediction, but also the broader patterns and context, leading to more effective and informed decision-making in real-world applications.

Chapter 6

General Discussion and Conclusions

This chapter summarises the key findings of the thesis and explores their design implications. It also provides a general discussion, addresses limitations, and suggests directions for future work.

6.1 Summary and Key Findings

Chapter 1 introduced the increasing integration of ML into diverse fields, enabled by advances in algorithms, computational power, and large data availability. Despite its technical depth, ML is increasingly available to users who are not trained to understand its inner workings. This means that human-AI interaction has become increasingly important to study, especially as ML models become more complex. This thesis evaluated interaction with AI systems by investigating how users perceive and engage with text classifiers, with emphasis on IML and XAI. The decision to use QDA as an application area was made to introduce ambiguity in the data, mirroring real-world scenarios where the ground truth is not always well-defined, and allowing for a more nuanced exploration of user interpretation and critical reflection.

Chapter 2 reviewed the current state of the HCI and AI literature on human-AI interaction, IML and XAI. The literature highlights several challenges in designing for human-AI interaction, particularly around capability uncertainty and output complexity, and thus calls for a deeper understanding of the human experience and psychology interacting with these systems. IML and XAI emerge as promis-

ing fields, where the first enables users to engage iteratively with models, and the second provides interpretable outputs that touch on aspects beyond transparency. Prior research has focused on guidelines and implications of IML, but a significant research gap remains in utilising it as a means to understand broader user interaction dynamics. On the other hand, user studies in XAI are considered essential for evaluating interpretability techniques, but user biases have been shown to affect the subjective evaluations that are commonly designed. This chapter motivated the research questions by identifying opportunities to explore user interactions with a functional IML tool and the effectiveness of XAI in text classification through model outcome prediction rather than self-reported metrics.

Chapter 3 presented a user study in which 20 participants who were not experts in ML used TACA, an IML tool designed and developed to assist users in QDA. Through a combination of a quantitative analysis of the automated interaction logs and a qualitative analysis of the responses in a semi-structured interview, the aim of the study was to answer the two RQs:

- **RQ1: How do non-expert users perceive ML when analysing ambiguous data?**
- **RQ2: How do non-expert users' perceptions of ML influence their interaction with it?**

The data collected revealed several key findings: 1) ML has value in the QDA workflow as the coding suggestions encouraged a more critical analysis of data, 2) ML encourages reflexivity when participants are confronted with conflicting classifications that challenge their own analysis, 3) users with no experience in ML tend to perceive the model as an external, objective source of advice, and 4) they hold themselves accountable when the model does not perform well, providing a variety of justifications regarding data set size, analysis structure, and manual mistakes.

As mentioned above, the participants in this study revealed a tendency to perceive the model as objective and authoritative, blaming themselves for the shortcomings in performance. An important and unexpected consequence is that the participants engaged only to a limited extent with the IML features of TACA. Building

on these findings, Chapter 4 reported on an autoethnography on TACA from the perspectives of developer, researcher and participant. The study aimed to uncover nuances in the interactions with the tool that might have been missed in the previous study due to several other limitations, such as the lack of familiarity with the data and the short-term, isolated nature of the study design. The study allowed for a direct comparison between the results of the qualitative analysis conducted on the interviews in the previous study and those of the analysis conducted with the use of TACA. This chapter also discussed the findings beyond the scope of QDA and thus attempted to answer the following RQs:

- **RQ3: How can IML be used to support the analysis of ambiguous data?**

The key findings of the study include: 1) as opposed to the experiences of the participants in the user study, the model was not viewed as an objective source of advice due to prior experience in ML and in development, 2) the iterative engagement with the model suggestions *and* the training data prompted reflections on potential structural changes to the analysis as well as uncovering additional insights, 3) balancing usability with model transparency could support both the efficiency and accuracy in IML by facilitating the identification of classification inconsistencies, and 4) it seems unrealistic to expect ML to generate radical new knowledge that can challenge existing assumptions.

Chapter 5 builds on the considerations on model transparency made in the previous chapters to evaluate the two most widely used local interpretability techniques, LIME and SHAP, and a novel global technique using LLM-generated summaries of LIME word importance weights. The study involved 128 participants who were asked to predict model behaviour based on the explanations, an approach aimed to address a methodological research gap of XAI user studies in text classification, in order to answer the following RQ:

RQ4: How do interpretability techniques affect users' ability to predict ML model behaviour?

Mirroring the design of a previous user study evaluating LRP saliency maps used in image classification, the question was split into 4 sub-questions:

- RQ4.1: Do SHAP and LIME generated word importance heat maps assist participants in predicting the outcome of a text classifier?
- RQ4.2: Are LLM-generated summaries of LIME word importance weights an effective interpretability technique?
- RQ4.3: What are the effects of interpretability techniques on the confidence of predictions of the model outcome?
- RQ4.4: How do different interpretability techniques influence users' attention towards specific features, and what effect does this have on their ability to understand overall model behaviour?

The data collected revealed the key findings: 1) none of the interpretability techniques had substantial effects on precision or confidence in outcome prediction, 2) participants with no explanations identified more topics, those exposed to LIME and SHAP heat maps identified more terms, and those exposed to LLM-generated summaries identified more features overall, and 3) the features mentioned by the participants were not indicative of model behaviour.

The contributions made by this thesis are summarised in Table 6.1.

6.2 Design Implications

The work presented in this thesis has several implications on the design of human-AI interaction with text classifiers, which are summarised in this section.

6.2.1 Chapter 3 Design Implications

- Systems used for analytical purposes should be designed to promote critical engagement with data by providing clear and insightful feedback on both generated classifications and manually labelled data samples to allow for comparisons in classification.
- Designers should account for the non-expert perception of the ML model as an external source of objective advice and the consequent self-blame for errors. Clearly communicating the probabilistic nature of ML and providing

Table 6.1: A summarised list of contributions made by this thesis.

Contribution Type	Contribution	Chapter
System	Designed, developed and implemented TACA, a fully-functional IML application to assist users in QDA.	Chapter 3, Chapter 4
Methodological	Conducted studies on ML using ambiguous data rather than well-defined ground truth, allowing for exploration of user interpretation and decision-making.	Chapter 3, Chapter 4, Chapter 5
Methodological	Conducted the first autoethnography on IML, demonstrating how self-study can generate new insights even after a system has been designed and evaluated.	Chapter 4
Methodological	Conducted the first user study based on model outcome prediction (forward simulation) in text classification in order to address methodological limitations of alternative approaches in XAI.	Chapter 5
Empirical	Identified the perceived role and value of ML in analytical processes.	Chapter 3, Chapter 4
Empirical	Identified the non-expert perception of ML as an objective source of advice and self-blame for poor model performance.	Chapter 3
Empirical	Reported on personal experiences with IML, noting the influence of prior ML knowledge, structural reflections on the analysis prompted by training and predicted data, the role of transparency, and limitations of ML to produce radically new insights.	Chapter 4
Empirical	Provided evidence of the lack of substantial effects on outcome prediction of the two most widely used local interpretability techniques, LIME and SHAP, and a proposed global technique based on LLM-generated summaries, in text classification.	Chapter 5
Empirical	Demonstrated that participants without explanations identify more topics, while those with LIME/SHAP focus on terms, and LLM-generated summaries lead to identifying more features overall, though these features do not reflect model behaviour.	Chapter 5

explanations for model behaviour could help users develop a more nuanced understanding and reduce unwarranted self-blame.

- ML applications can benefit from data aggregation to support model inspection and feedback assignment. However, designers should include complementary features, such as tooltips, additional context, and interactive tutorials, to help users understand the relationship between grouped data and re-training the model.

6.2.2 Chapter 4 Design Implications

- The design of IML systems should account for different user needs and offer flexibility for those who prioritise efficiency and accuracy differently based on their constraints and goals. Users constrained by time or resources may benefit from batch re-labeling for quicker progress, while those able to invest more effort can enhance model performance through meticulous, individual refinements.
- In IML, interpretability should be designed to take an additional role by optimising the user’s contribution to the iterative learning process of the model.
- Designers should incorporate tools that encourage users to reflect on and evaluate the ground truth data, helping non-experts calibrate trust by understanding model outputs as data-dependent, and assisting practitioners in identifying labeling inconsistencies to improve model performance before hyperparameter tuning.
- Tools using ML to assist with analytical processes should be designed with the assumption that users will still need to make critical decisions, as ML models only reflect patterns from training data and rely on human insight for deeper and contextually grounded interpretations. This approach ensures that users view model output as prompts for further analysis rather than as conclusive insights.

6.2.3 Chapter 5 Design Implications

- Interpretability techniques need to be tailored to specific tasks depending on the number of identifiable features, because techniques that work well on fewer and more tangible features like in images may not translate effectively to tasks with more abstract features, such as text.
- Interpretability techniques should not only indicate important features, but also encourage users to consider broader context and patterns. Tools could combine local and global explanations to help users develop holistic understanding of the decision-process of the model.
- Techniques should be evaluated also on task performance rather than relying only on subjective measures such as perceived usefulness, as most participants reported finding the heat maps useful despite not actually making more accurate predictions.

6.3 Discussion

The findings reported in this thesis confirm the importance of research in human-AI interaction. The following paragraphs provide an overview of how these findings can be explained and what they mean in a more general context.

6.3.1 Extended Benefits of AI Beyond Model Performance

Chapters 3 and 4 revealed that there is value in applying ML to the analysis of ambiguous data. In this context, the benefits of ML extend beyond the automation or acceleration of data analysis that can be found in other domains, such as finance, manufacturing, and recognition tasks. When users are presented with conflicting suggestions from the model, their own interpretation of the data is questioned. Consequently, users are encouraged to challenge their positionality and critically reflect on their initial assumptions. In the analytical process, this reflexivity is crucial, as it allows for the adjustment of potential biases, leading to more nuanced and contextually aware interpretations.

AI systems are often designed for accuracy with specific emphasis on performance measures, ensuring that, once the model is deployed, it can generalise learned patterns to unseen data. The findings of Chapters 3 and 4 suggest that this approach is only appropriate in domains where a well-defined ground truth exists and accuracy is essential for practical application. Achieving an initial average of 0.58 for the F1 score in the study of Chapter 3, the model deployed in TACA was not particularly accurate. However, users were still able to derive meaningful insights and engage in critical reflection from the model output. This is because, as long as the initial interpretations are challenged, it matters less whether the classifications of the model are strictly accurate.

The utility of AI systems extending beyond performance has implications on a wide range of applications. For example, in educational technology, a system might suggest learning paths or highlight areas of struggle for students. Even if the classification of strengths and weaknesses is not perfect, the system can still encourage teachers to reflect on their students' needs and to adapt teaching methods. In more creative domains, the system could suggest novel ideas, propose stylistic adjustments, or highlight potential areas for improvement. When suggestions do not perfectly align with the user's intentions, they can still encourage users to reconsider their approach or explore alternative perspectives, enriching the final work through critical reflection.

In general, these findings align with calls for a shift in how AI is evaluated in domains where insights and reflections matter. As found in Chapter 3 and discussed in Chapter 4, AI system designers can benefit from prioritising transparency, flexibility and engagement over just performance. Gillies et al., 2016 argued that popular ML approaches do not fully exploit the nature of applied ML as a co-adaptive process. Indeed, not only does the user influence the behaviour of the model, but they also adapt to using the system more effectively and may even modify their goals based on what is learned through the tool. The intrinsic value of AI in these cases lies therefore in its ability to support an evolving understanding.

6.3.2 Challenges in Perceiving AI as an Objective Authority

Despite the advantages of using AI to promote critical reflection, Chapter 3 also revealed a potential challenge: participants who were non-expert in ML tend to perceive the model as an external, objective source of authority. This perception can lead users to overly rely on the output of the model, viewing it as more factual or unbiased than it actually is. While challenging personal interpretations is beneficial, it arguably requires a balance; if users fail to acknowledge the subjectivity inherent to the data and see the suggestions of the model as objective truths, the intended reflexivity can be undermined. Not only does this misinterpretation limit the engagement with the model (interrupting the IML cycle), it also introduces the risk of potentially reinforcing existing biases rather than revealing them.

There are many cases in which the application of AI to ambiguous data can pose significant risks. AI is currently being utilised in various aspects of hiring and recruitment (Hunkenschroer and Kriebitz, 2023). When assessing candidate qualities like “potential for personal growth”, ML models can present recommendations that are interpreted as definite, resulting in biased hiring decisions. Automated content moderation on social media relies on context-dependent and culturally nuanced definitions of offensive or inappropriate content (Gorwa, Binns, and Katzenbach, 2020). If moderation relies excessively on AI without human interpretation, this poses the risk of unjust penalties and stifling diverse viewpoints, and potentially causing reputational damage to the platforms. In legal document review, AI is used to highlight relevant cases or suggest interpretations (Ashley, 2017), but non-expert users may risk missing crucial legal nuances, potentially leading to misinterpretation and harm their clients’ cases.

The tendency to view AI outputs as authoritative can perhaps be explained by considering several factors. ML has recently achieved remarkable results in natural language understanding, image recognition, and predictive modelling. Particularly influential are applications of generative AI, such as ChatGPT and Stable Diffusion, which have reshaped the public’s perception by providing tangible examples of AI producing human-like outputs. This rapid progress has been unprecedented and

has contributed to a perception of AI systems as highly accurate. In recent years, extensive media coverage and hype promoted by the tech industry has probably also contributed to heightened public perceptions of AI's capabilities and impact, often skewing these perceptions towards either an overly optimistic or pessimistic outlook on AI. This tendency towards polarised views, as noted by Brauner et al., 2023, can affect public expectations and potentially create unrealistic or fear-driven attitudes toward AI's role in society. In both cases, these perceptions are fuelled by the notion that AI is inherently high-performing and accurate, regardless of the complexities and potential biases in its underlying data and algorithms.

In reality, ML models are limited to capturing patterns present in the data they are trained on, and are therefore inherently shaped by any biases or limitations within that data. Chapter 4 provided a first-hand account illustrating how awareness of the model's dependence on training data helped calibrate perceptions of its reliability. However, findings in Chapter 3 showed that, while the interviews revealed that the non-expert participants shared this awareness, the same interviews, as well as their behaviour and experience with TACA, demonstrated that this awareness alone was not enough to prevent them from interpreting the outputs as objective and unbiased conclusions. This suggests that more guidance is needed to help non-expert users align their perceptions more closely towards the probabilistic nature of ML models and their inner workings. Performance metrics and confidence scores alone can be confusing or misleading, and non-expert users may not have the technical knowledge to interpret ML-specific measures. In IML, they tend to favour model quality as the task completion criteria (Dudley and Kristensson, 2018; Q. Yang, Suh, et al., 2018), which is undesirable when the data is ambiguous, as discussed in Chapter 4 (see Sections 4.5.3.1 and 4.6.1.1).

The concept of guidance often emerges in the literature on "machine teaching", a term that has been recently been reframed by Simard et al., 2017 to place focus on the efficacy of the teaching process in IML by measuring performance relative to human costs, such as productivity, interpretability, and robustness. In a study exploring how novice users trained an IML model using a web-based sketch recog-

nition tool, Sanchez et al., 2021 concluded that providing structured guidance in sequencing examples and explaining model feedback could help novice users build more effective training strategies, enhancing their engagement and improving their understanding of the model. The studies reported in Chapters 3 and 4 similarly emphasise the need for guidance to support non-expert interactions and drive model refinement.

These observations point to model transparency as a crucial element in effectively managing user expectations and fostering more appropriate reliance on AI systems. In fact, studies have shown that enhancing transparency by providing users with explanations of model behaviour and limitations can help non-experts better understand model outputs (Poursabzi-Sangdeh, Daniel G. Goldstein, et al., 2021b). The benefits of interpretability would therefore appropriately address the findings of Chapter 3 by mitigating the risk of overestimating the objectivity of model outputs. In IML, transparency also promotes a collaborative relationship between the user and the model, where the user feels encouraged to question and refine their interpretations rather than accepting the output as final (Stumpf, Rajaram, Li, Wong, et al., 2009; Amershi, Cakmak, et al., 2014). However, the reported advantages of introducing model transparency are highly dependent on the choice and quality of the interpretability techniques used, as unsuitable or poorly tailored explanations can still lead to confusion or reinforce misconceptions (Nourani et al., 2019). In a study from Bansal et al., 2021, explanations were actually found to *increase* the chance that users accept the AI's recommendation, regardless of its correctness. The evaluation of these techniques is therefore essential to ensure that they effectively help users develop an accurate and balanced understanding of AI outputs and limitations.

6.3.3 Model Transparency to Manage User Perceptions of ML

The study presented in Chapter 5 was conducted in isolation, with the intention of building on the findings when integrating interpretability techniques into TACA in the future. This approach was necessary due to the lack of previous work specifically evaluating techniques in text classification following task performance mea-

sures such as the “forward simulation” approach (Belle and Papantonis, 2021). As discussed in Sections 2.3.2 and 5.3.1.1, various evaluation methods have revealed contrasting results both between and within ML application domains. Even considering different evaluation methods, very few studies have been conducted on text classification compared to image recognition, recommender systems, and decision-support systems.

Although the particular task of predicting how a piece of text would be classified might seem trivial, the insights gained from this task become important when considering the broader implications. When techniques enable model behaviour prediction, users are more likely to identify inconsistencies, question automated decisions, and adjust their reliance on the model based on a deeper understanding of the outputs. The two main challenges in designing human-AI interaction are capability uncertainty and output complexity (Q. Yang, Steinfeld, et al., 2020). If interpretability techniques help users anticipate model behaviour, they can mitigate both of these issues by fostering a more transparent and reliable interaction with the AI system.

The fact that the two most widely used model-agnostic interpretability techniques were insufficient for participants to anticipate model behaviour reflects the finding of other evaluations of different techniques in different application domains (Buçinca, Lin, et al., 2020; Waa et al., 2021). The use of ambiguous data introduced an additional layer of complexity in the study that required participants to actively interpret the labelling patterns and the output of the model rather than relying on straightforward predictions. It is easy to imagine how a different data set would have simplified the predictive task. However, ambiguity mirrors many real-world scenarios where data often lacks clear boundaries or universally agreed-upon interpretations. These are the scenarios around which the studies in Chapters 3 and 4 were designed, and arguably where transparency is needed the most to understand model behaviour. This is because the user is also required to recognise and understand the subjective perspective embedded in the training dataset by the manual labeller, and that is reflected in the model output. The example given in

Section 5.3.1.1, where text samples related to price were manually labelled in the category “place”, illustrates how introducing subjectivity in the data set complicates the interpretation of model behaviour. In this case, the participants might have benefitted from techniques that paid more attention to the underlying labelling assumptions and highlighted potential ambiguities.

Based on the findings reported in Chapter 5, it seems unreasonable to expect that the evaluated interpretability techniques could provide meaningful value in an IML system like TACA. Instead, it is possible to draw on previous work in XAI to speculate that a combination of local and global techniques or an interactive approach could perhaps yield more promising results (B. Kim, Khanna, and Koyejo, 2016; Scott M. Lundberg and S.-I. Lee, 2017; B. Kim, Wattenberg, et al., 2018; Waa et al., 2021). This is also suggested by the findings on feature attention reported in Section 5.3.2: local explanations drew too much attention to individual terms, whereas the global explanation based on LLM-generated summaries provided an overview that was too general and abstract compared to the given tasks. At first glance, these two approaches seem to complement each other and could be combined to balance detailed insights with broader context.

In general, the vast range of evaluation methods used to measure the effectiveness of interpretability techniques is a testament to the difficulty of establishing a universal standard. In a systematic review of the evaluation methods used in XAI, Nauta et al., 2023 identified 12 different measurable explanation quality properties. In this context, it seems unreasonable to design techniques that excel across all properties simultaneously. A more practical approach could be to tailor or choose techniques based on the application in which they are used. The need for a context-specific approach is underscored by the findings of the study presented in Chapter 5, which diverged from those of similar work on image classification (Alqaraawi et al., 2020).

6.4 Limitations

The studies presented in Chapters 3, 4 and 5 confirm that not only is designing for human-AI interaction challenging, but so is the design of user studies to investigate these interactions. In “interaction”, there are multiple dimensions to consider, including user expectations, mental models, behaviours, and decision-making, as well as the interpretability, influence, and perceived role of the ML model. The interplay of these factors creates a vast space for research that requires an interdisciplinary perspective to capture the nuanced dynamics between human cognition and AI.

There are many aspects of interaction that the thesis did not evaluate due to the scope of the research. User trust stands out as a key factor in people’s interactions with AI-infused systems and an aspect that the literature on human-AI interaction has frequently focused on (Jacovi et al., 2021; Ueno et al., 2022; Hyesun, Prabu, and Arun, 2023; Milana, Costanza, and Fischer, 2023). Defined by J. D. Lee and See, 2004 as “an attitude that an agent will achieve an individual’s goal in a situation characterised by uncertainty and vulnerability”, trust in AI becomes crucial in high-stakes scenarios such as healthcare, finance, and criminal justice. In a survey on empirical methodologies used to evaluate trust in AI-assisted decision-making, Vereschak, Bailly, and Caramiaux, 2021 specifically recommend using established questionnaires that comprise all the key elements defining trust and designing studies to include interactions over a long period of time to measure how trust can be developed, damaged or repaired.

Although the studies were not designed to address it directly, trust is an underlying theme of this thesis, taking different forms in Chapters 3, 4 and 5. In Chapter 3, the perception of the model as being objective and authoritative echoes previous research on non-experts in ML to trust algorithmic outputs, even when they lack transparency or interpretability (Q. Yang, Suh, et al., 2018). In contrast, Chapter 4 illustrates how knowledge of the inner workings can mediate reliance by enabling users to more critically assess model outputs and adjust their trust accordingly. The findings in Chapter 5 suggest that merely presenting explanations does not guarantee transparency, as effective transparency should enable users to

anticipate system behaviour. Consequently, trust in the model may not have been meaningfully calibrated, as insufficiently informative explanations fail to improve appropriate reliance on AI (Buçinca, Malaya, and Gajos, 2021).

The focus of the work presented was also necessarily narrowed down to a specific application area: text classification. However, AI is widely used in other domains too, such as image recognition, recommendation systems, and generative applications, each of which presenting their own unique challenges for human-AI interaction. Chapter 5 is an example of how the same methodology applied in one domain revealed different findings in another: saliency maps had a significant effect in model outcome prediction in image classification but did not have the same effect in text classification. This shows that the findings reported in this thesis are not automatically generalisable to *all* applications of AI, and should instead be investigated in other domains to confirm or challenge their applicability across contexts.

The difficulty in designing and running the studies reported was the reason for various trade-offs and compromises that were made to balance ecological validity with practical feasibility. Participant recruitment posed a serious challenge in conducting the study reported in Chapter 3. Eligibility requirements included: experience in QDA, lack of experience in AI/ML, having an appropriately long data set pre-analysed using compatible QDA software, and a computer with an operating system version compatible with TACA. Additionally, some participants encountered various errors during the initial installation and setup phase of TACA. These issues, which stemmed from initial configuration challenges, were difficult to replicate and resolve without access to the participants' data sets, which could not be shared due to ethical restrictions. For this reason, the requirement for a pre-analysed data set was eventually dropped and participants without a data set were given a collection of newspaper restaurant reviews.

The decision to use restaurant reviews introduced limitations to the study that partly motivated the autoethnography reported in Chapter 4. Namely, that the majority of the participants lacked deep familiarity with the data set, or at least the same level of contextual understanding as those who used their own data. Addition-

ally, the analysis of the restaurant reviews was not conducted out of genuine interest by this group of participants, unlike the analysis of participants using their own data set, which would likely have held more personal relevance and meaning. The lack of intrinsic motivation may have influenced both the depth and engagement of the participants' analysis, while the unfamiliarity with the data may have influenced the participants in the review of the outcome of the classifier.

Chapters 3, 4 and 5 all present studies based on the same XGBoost classifier. Although Gradient Boosting models are effective and still used for many classification and regression tasks, NLP has recently been revolutionised by the Transformer architecture, which is far more commonly used today. The three studies conducted were designed around classifications that were conflicting with the training data (false positives and false negatives), since these could create situations where the user would need to critically evaluate and interpret the output of the model. To obtain a sufficient number of false positive and false negative classifications, it would have been necessary to either artificially limit the performance of the model or significantly increase the size of the data sets. The first option would have arguably compromised ecological validity, while the second would have reduced practical feasibility. Nonetheless, using a Transformer would have revealed findings that would have been perhaps more applicable to real-world modern applications. For example, the study reported in Chapter 5 evaluated strictly model-agnostic interpretability techniques since Gradient Boosting models are gradually being replaced by Transformers in many applications, but the use of Transformers would have allowed for the evaluation of model-specific techniques, such as Integrated Gradients (Sanyal and Ren, 2021). This would have yielded results that are more useful for current applications of ML in text.

6.5 Future Work

The thesis revealed several potential avenues for future work, some of which build on the findings reported, while others could address the limitations discussed.

Chapter 3 reported the perception of the ML model as an external and objective

source of advice of non-expert users, who considered misclassifications as suggestions and rarely questioned the performance of the model. One could argue that this perception was due to the ambiguity of the data more so than a bias towards AI. Although the participants in the interview specifically used terms like “objective” and almost always blamed themselves when encountering unexpected outputs, future work could explore this in more depth. For example, a between-subject user study could be designed where one group of participants is told that the coding suggestions are generated by AI, while another is told that these belong to a human coder. An analysis of the responses in a post-study interview would then confirm or reject this finding.

Chapter 4 reported on personal experiences regarding the use of TACA as an IML tool to assist in the coding phase of thematic analysis. One aspect that emerged is the potential benefit of using interpretability techniques not just for passive inspection of model behaviour, but also to drive model feedback within the IML cycle. Recent work has confirmed that interpretability techniques can significantly enhance the user’s ability to correct the model by providing explanations that highlight misclassifications and areas of improvement (Teso et al., 2023). Implementing explanations in TACA could reveal complementary results by evaluating how model transparency affects non-expert perceptions of the model, and whether these are still effective in driving model behaviour when the ground truth is ambiguous.

Chapter 5 concluded that the interpretability techniques LIME and SHAP do not have any substantial effects on model outcome prediction. This was partly explained by the fact that the samples shown with the heat maps as example classifications contained different features than the ones present in the task samples (see Appendix C.7). Future work could confirm this through a think-aloud protocol using the same experimental design. During the user study reported, this approach was attempted in pilot studies involving two participants at a time so that they could discuss salient feature in order to come to the same decision on model outcome. The study design was eventually discarded due to technical and recruitment challenges. However, re-introducing a think-aloud protocol could reveal the decision-making

process of users exposed to interpretability techniques and thus provide an explanation for the lack of effects.

There is also a clear opportunity to repeat the three studies with a Transformer model. Taking advantage of the system contribution in TACA, it could be possible to simply replace the XGBoost model with a BERT model (Koroteev, 2021) fine-tuned for text classification. Similarly, BERT could be used to generate the classifications for the study evaluating interpretability techniques. Instead of using pre-generated GloVe word embeddings, the BERT model would dynamically generate contextual embeddings, meaning that, when sampling example sentences based on vector distance from the task sentences, it is more likely that the example sentences are closer in meaning to the task sample. This would evaluate in more depth whether the features of example sentences in local explanations are indeed representative enough to support a meaningful understanding of model behaviour, potentially confirming or refuting the implication that current examples lack sufficient representativeness.

The three studies reported in this thesis were designed by deliberately introducing ambiguous data. Future work could explore different application areas of ML where this is also true, such as legal and judicial analysis or human behaviour predictions. This would be particularly useful considering that the eligibility requirement of experience in QDA meant that most of the participants recruited for the study reported in Chapter 3 had a background in HCI or psychology (see Table 3.1). Drawing from a different participant pool is important because it would provide insights into how individuals with different background and varying expertise interpret the data and the output of the model. For example, the participants in the user study easily understood the terms of the Confusion Tables (true positive, false negative, etc.), but this could be due to their familiarity scientific terminology and experience in data analysis. Different backgrounds could reveal alternative perspectives and provide a more comprehensive understanding of user needs in diverse application areas.

Finally, future work could also frame evaluations of IML systems and inter-

prebability techniques within user trust. The literature on human-AI trust has two main areas of interest: defining what trust is and what factors affect it (Vereschak, Alizadeh, et al., 2024). Additional user evaluations on TACA would thus fall in the second camp by specifically evaluating the main factors of trust in AI identified by Glikson and Woolley, 2020: tangibility, transparency, performance, task characteristics, anthropomorphism, and socially-oriented behaviours. Studies could employ the practical guidelines for the methodology of studying trust in decision-making AI applications identified by Vereschak, Bailly, and Caramiaux, 2021. Moreover, when measuring trust in AI, it can be beneficial for future work to design user studies that incorporate elements of risk, as this aligns with the definition of trust by J. D. Lee and See, 2004 concerning “a situation characterised by uncertainty and vulnerability”, and simulates real-world decision-making contexts where trust is most relevant (Milana, Costanza, and Fischer, 2023).

6.6 Conclusion

This thesis reports on a series of studies aimed at evaluating the interaction with ML text classifiers and interpretability techniques. Building on previous work, the thesis reiterates the importance of studying human-AI interaction as ML continues to be implemented in a growing number of application areas with systems that do not require technical knowledge from the users.

In addition to a system contribution in a fully-functional IML application to assist in thematic analysis, this thesis combines quantitative, qualitative, and self-study methods, contributing to both methodological and empirical insights by revealing the following key findings:

- In the absence of a well-defined ground truth, participants who were non-expert in ML were subject to perceiving ML outputs as objective, often attributing poor performance to themselves rather than questioning the model.
- ML can play an important role in supporting analytical processes, although users should be aware that models reinforce existing patterns in the training data and do not generate radical new knowledge.

- The most commonly used local interpretability techniques, LIME and SHAP, had no impact on participants' ability to predict model outcomes in text classification tasks.
- Different interpretability techniques shifted the participants' focus onto different features, but these features rarely aligned with the actual behaviour of the model.

These findings were possible due to the use of training data sets with an ambiguous ground truth, reflecting real-world complexities where labels are often subject to interpretation. Unlike ML classification tasks typically used in research that use clearly defined labels, this approach provided a more realistic assessment of user behaviour, capturing how subjectivity affects the perception of ML. Hopefully, this work will inspire future researchers to explore similar scenarios in human-AI interaction and draw attention to the unique perspective of HCI research on AI.

Appendix A

Chapter 3 Supplementary Material

A.1 User Study Participant Information Sheet (Participants' Own Data)

**UCL DIVISION OF PSYCHOLOGY
AND LANGUAGE SCIENCES**

Information Sheet for Participants in Research Studies

You will be given a copy of this information sheet.

Title of Project: **Thematic Analysis Assistant Tool: Interacting with Machine Learning** This study

has been approved by the UCL Interaction Centre (UCLIC) Research Department's Ethics Chair Name,

Address and Contact Details of Investigators:

Principal

Dr. Enrico Costanza

investigator:

UCL Gower Street, London, UK
+44 (0)20 7679 7181
e.costanza@ucl.ac.uk

Co-investigators: Prof. Mirco Musolesi

UCL Gower Street

London, UK

m.musolesi@ucl.ac.uk

Federico Milana

UCL Gower Street

London, UK

federico.milana.18@ucl.ac.uk

The **aim of this research** is to explore and evaluate interactions with a machine learning model. As a result, we present an interactive desktop application designed to assist the coding phase in thematic analysis by classifying new sentences into user-defined themes.

The study involves the use of an **interactive desktop application and a follow-up remote interview**. The use of the application involves importing an interview transcript that has already been coded by the participant either in Microsoft Word, NVivo, MAXQDA or Dedoose. The application will classify new sentences into user-defined themes and display various tabs: 1) "Text", where user-coded and classified sentences are highlighted within the transcript, 2) "Codes", where the code-theme lookup table is displayed, 3) "Keywords", where sentences are grouped into the most frequent keywords for each theme, and 4) "Confusion Tables", a table version of confusion matrices for each theme. Words in "Codes" and "Confusion Tables" can be clicked to view sentences, while keywords and single sentences in "Keywords" can also be dragged to different columns (or bin) to retrain the model. The expectation is that setting up your files and the tool should take about 20 minutes, interacting with the application 20 minutes, and the follow-up interview 20 minutes.

You are free to stop the study at any point, without needing to explain why. There are no particular risks associated with your participation other than those associated with the use of standard computer equipment.

Exclusion Criteria. To take part, you must have an interview transcript .docx file, or several transcripts merged in a single file. The transcript must either be coded in Microsoft Word (with comments), NVivo, MAXQDA, or Dedoose.

Data and Information. All data will be handled according to the General Data Protection Act 2018 and will be kept anonymous. We will audio-record the interview, transcribe these recordings, and take notes; the audio recordings will be stored on a secure server and then deleted once the data analysis process is complete. The data and the results of the analysis may be made publicly available only in aggregate or anonymous form (i.e., not revealing your identity).

Any information that is obtained in connection with this study and that can be identified with you will remain confidential and will be disclosed only with your permission or as required by law. The information collected during the experiment will be kept separately from your personal identity. If you decide to withdraw from the

study after taking part in the interview, due to the anonymity of the data, it may not be possible to identify and destroy it.

Concerns or complaints.

Should you have any concern or complaint, you can contact us at any point via email (Federico Milana federico.milana.18@ucl.ac.uk or Enrico Costanza e.costanza@ucl.ac.uk).

A.2 User Study Participant Information Sheet (Newspaper Restaurant Reviews)

UCL DIVISION OF PSYCHOLOGY
AND LANGUAGE SCIENCES

Information Sheet for Participants in Research Studies

You will be given a copy of this information sheet.

Title of Project: **Thematic Analysis Assistant Tool: Interacting with Machine Learning** This study

has been approved by the UCL Interaction Centre (UCLIC) Research Department's Ethics Chair Name,

Address and Contact Details of Investigators:

Principal

Dr. Enrico Costanza

investigator:

UCL Gower Street, London, UK
+44 (0)20 7679 7181
e.costanza@ucl.ac.uk

Co-investigators: Prof. Mirco Musolesi

UCL Gower Street

London, UK

m.musolesi@ucl.ac.uk

Federico Milana

UCL Gower Street

London, UK

federico.milana.18@ucl.ac.uk

The aim of this research is to explore and evaluate interactions with a machine learning model. As a result, we present an interactive desktop application designed to assist the coding phase in thematic analysis by classifying new sentences into user-defined themes.

The study involves using thematic analysis to code 21 newspaper restaurant reviews into 4-8 themes, using an interactive desktop application, and taking part in an interview in a remote meeting. The use of the application involves importing the reviews coded either in Microsoft Word (using comments), NVivo, MAXQDA or Dedoose. The application will classify new sentences into the user-defined themes and display various tabs: 1) "Text", where user-coded and classified sentences are highlighted within the transcript, 2) "Codes", where the code-theme lookup table is displayed, 3) "Keywords", where sentences are grouped into the most frequent keywords for each theme, and 4) "Confusion Tables", a table version of confusion matrices for each theme. Words in "Codes" and "Confusion Tables" can be clicked to view sentences, while keywords and single sentences in "Keywords" can also be dragged to different columns (or bin) to retrain the model. The expectation is that coding the restaurant reviews should take 2 to 3 hours, using the tool should take about 30 minutes, and the interview should last about 30 minutes.

You are free to stop the study at any point, without needing to explain why. There are no particular risks associated with your participation other than those associated with the use of standard computer equipment.

Exclusion Criteria. To take part, you must have experience in qualitative analysis, and no experience in artificial intelligence/machine learning. The restaurant reviews must either be coded in Microsoft Word (using comments), NVivo, MAXQDA, or Dedoose.

Data and Information. All data will be handled according to the General Data Protection Act 2018 and will be kept anonymous. We will audio-record the interview, transcribe these recordings, and take notes; the audio recordings will be stored on a secure server and then deleted once the data analysis process is complete. The data and the results of the analysis may be made publicly available only in aggregate or anonymous form (i.e., not revealing your identity).

Any information that is obtained in connection with this study and that can be identified with you will remain confidential and will be disclosed only with your permission or as required by law. The information collected

during the experiment will be kept separately from your personal identity. If you decide to withdraw from the study after taking part in the interview, due to the anonymity of the data, it may not be possible to identify and destroy it.

Concerns or complaints.

Should you have any concern or complaint, you can contact us at any point via email (Federico Milana federico.milana.18@ucl.ac.uk or Enrico Costanza e.costanza@ucl.ac.uk).

A.3 User Study Informed Consent Form

CONSENT FORM

Please complete this form after you have read the Information Sheet explaining the research.

If you have any questions arising from the Information Sheet or explanation already given to you, please ask the researcher before you decide whether to join in. You will be given a copy of this Consent Form to keep and refer to at any time.

This study has been approved by the UCL Interaction Centre (UCLIC) Research Department's Ethics Chair, Project ID No: UCLIC_2022_004_costanza

Department: UCL Interaction Centre

Name and Contact Details of the Principal Researcher:	Name and Contact Details of the Researchers:
Prof. Enrico Costanza, e.costanza@ucl.ac.uk	Federico Milana, federico.milana.18@ucl.ac.uk Prof. Mirco Musolesi, m.musolesi@ucl.ac.uk Dr. Amid Ayobi, amid.ayobi@ucl.ac.uk

Name and Contact Details of the UCL Data Protection Officer: Nadia Berthouze, data-protection@ucl.ac.uk

I confirm that I understand that by ticking each box below I am consenting to this element of the study. I understand that it will be assumed that unticked boxes means that I DO NOT consent to that part of the study. I understand that by not giving consent for any one element that I may be deemed ineligible for the study.

1.		Tick Box
2.	I confirm that I have read and understood the Information Sheet for the above study. I have had an opportunity to consider the information and what will be expected of me. I have also had the opportunity to ask questions which have been answered to my satisfaction and would like to take part in the observation study.	
3.	I consent to participate in the study. I understand that my personal information, in the form of this consent form, will be used for the purposes explained to me. I understand that according to data protection legislation, 'public task' will be the lawful basis for processing.	
4.	I understand that all personal information will remain confidential and that all efforts will be made to ensure I cannot be identified.	
5.	I understand that my data gathered in this study will be stored anonymously and securely. It will not be possible to identify me in any publications.	
6.	I understand that my information may be subject to review by responsible individuals from the University for monitoring and audit purposes.	
7.	I understand the potential risks of participating and the support that will be available to me should I become distressed during the course of the research.	
8.	I understand the direct/indirect benefits of participating. In particular, I understand that I will receive £45.00 in compensation for participating.	
9.	I agree that my anonymized research data may be used by others for future research. (No one will be able to identify you when this data is shared).	
10.	I understand that the information I have submitted will be published as a report and I wish to receive a copy of it. Yes/No (delete as appropriate). If yes, please include your email address here: _____	Yes /No
11.	I hereby confirm that: (a) I understand the exclusion criteria as detailed in the Information Sheet and explained to me by the researcher; and (b) I do not fall under the exclusion criteria.	
12.	I am aware of who I should contact if I wish to lodge a complaint.	
13.	I agree for the data specified in the information sheet to be collected and stored on a secure server, and for an aggregate or anonymous version of this data to be made publicly available.	

If you would like to be contacted in the future by UCL researchers about participating in follow up studies to this project, or in future studies of a similar nature, please tick the appropriate box below.

Yes, I would be happy to be contacted in this way If yes, please include your email address here: _____	
No, I would not like to be contacted	

Name of participant

Date

Signature

Researcher

Date

Signature

A.4 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.

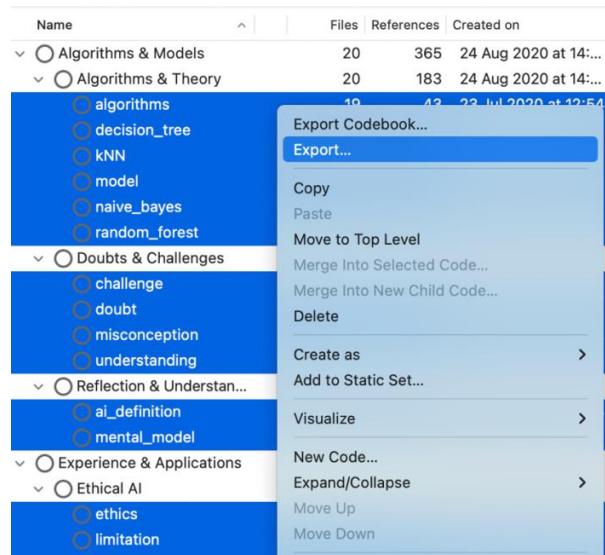
Setup

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:

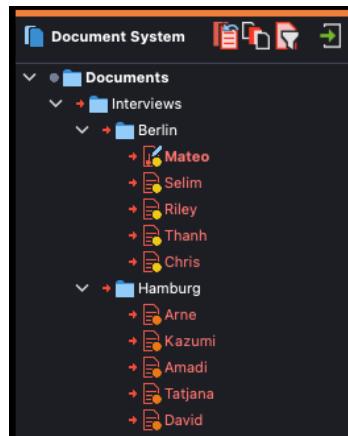
1. Inside NVivo, select all the codes at the lowest level -> right click -> Export...
(to quickly select all the codes: Ctrl/⌘ + A -> Ctrl/⌘ + click to deselect higher level codes)



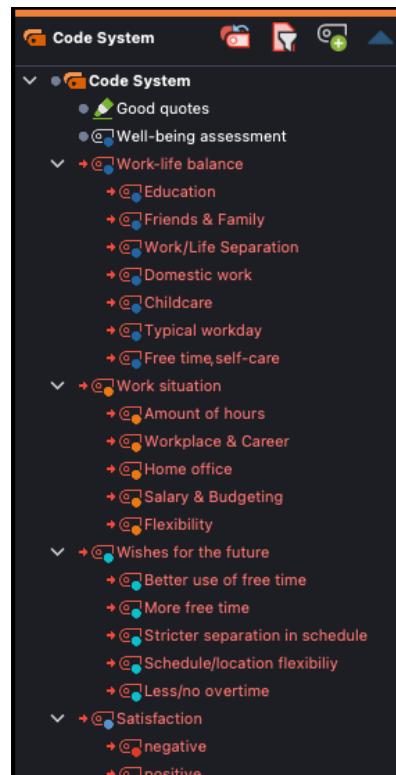
2. On Windows, after you click on “Export...”, select “Reference View”, the “Name” checkbox, and “Folder and Hierarchical Name” from the dropdown list.
3. Export all the codes in a separate folder. The folder should only contain .docx files, one for each code.

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

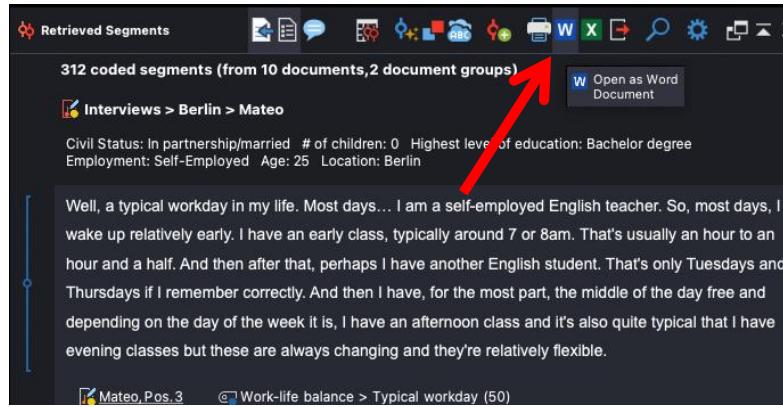
1. Inside MAXQDA, select the transcripts from the Document System pane:



2. Select the codes from the Code System pane:

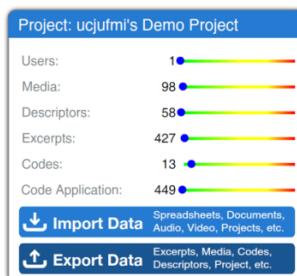


3. Open the Retrieved Segments pane and click on the **W** button to export all segments in a .docx file.



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

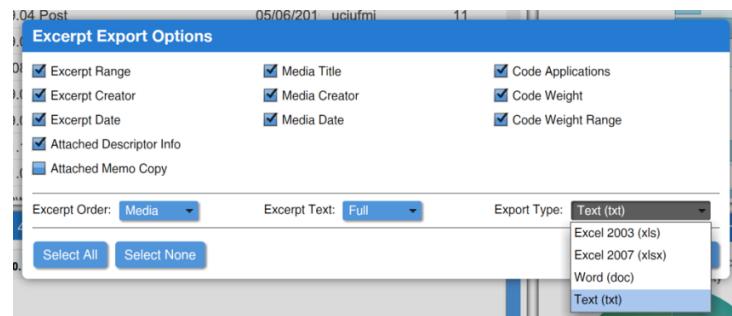
1. Inside Dedoose, click Export Data in the Project pane:



2. Click Export Excerpts in the popup:



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



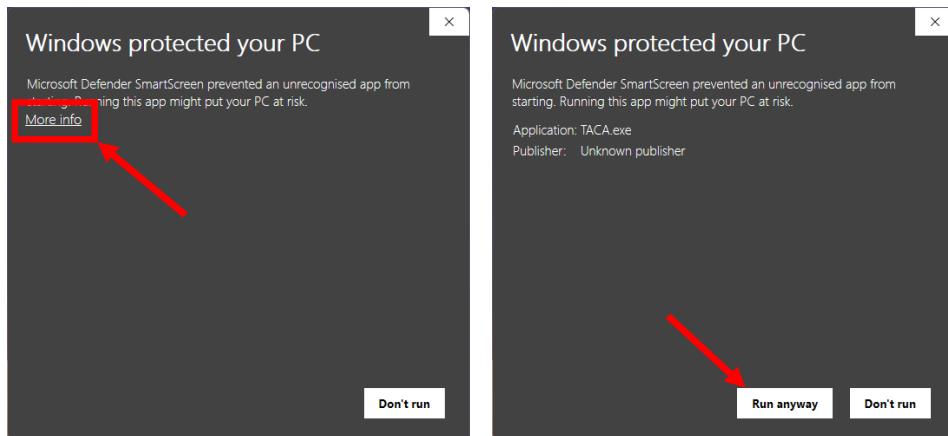
Step 2: Run the tool

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:



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

The Text tab displays a transcript of a conversation. Sentences are color-coded: grey for manually coded by the user, blue for coded by the model, and black for uncoded. A sidebar on the right lists themes with their respective counts.

Theme	Count
practices	496
study vs product	292
system perception	228
system use	715
value judgements	224

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

The Codes tab shows a table of imported codes. Each row includes the code name, count, and a tooltip containing examples of sentences manually coded with that code.

Code	Count
practices (496)	496
social (128)	128
study vs product (292)	292
system perception (228)	228
system use (715)	715
value judgements (224)	224

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

Text	Codes	Keywords ▾	Confusion 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)			every (11)	need (93)	email (15)
buy (30)			one (10)	say (92)	good (14)
eggs (26)			ate (10)	want (87)	time (14)
supermarket (25)			ent (10)	remember (80)	delivery (13)
make (24)			people (10)	delivery (79)	thought (12)
say (23)			do (9)	text (72)	website (12)
delivery (23)			bought (9)	quite (72)	say (11)
system (22)			the (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)	

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

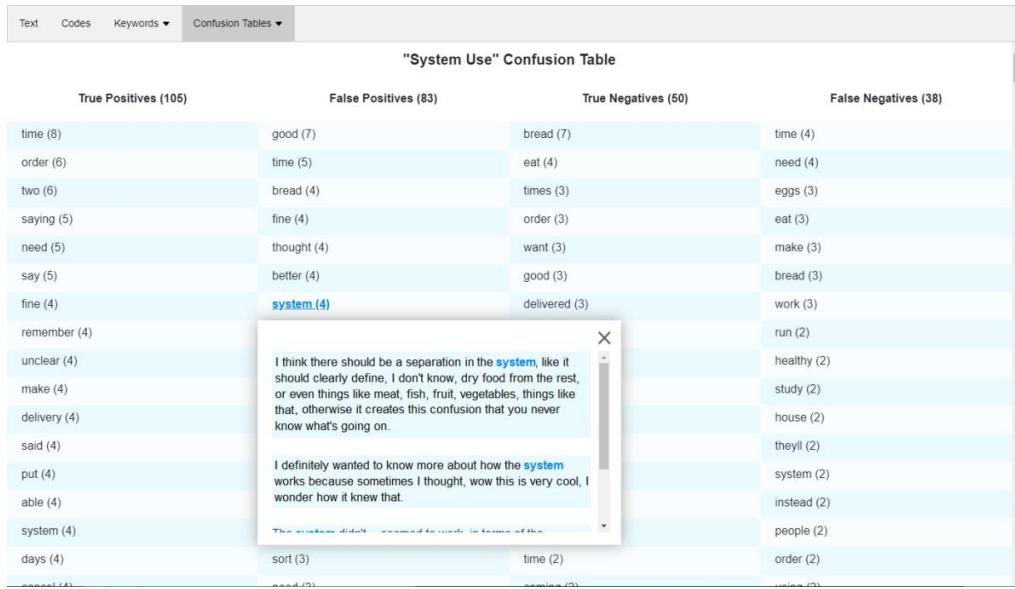
Here, the most frequent words are shown under each theme, along with a counter indicating the number of sentences that contain them. You can click on a word to reveal these sentences in a tooltip. Sentences coded by the model have a **blue** background, while those manually coded have a **grey** background.

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



The screenshot shows a 'Confusion Tables' tab in a software interface. The main table is titled '"System Use" Confusion Table' and has four columns: True Positives (105), False Positives (83), True Negatives (50), and False Negatives (38). The rows list words and their counts. A tooltip for the word 'system' (4) appears over the second column, containing two sentences from the dataset:

- 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.

A.5 Semi-structured Interview Script

Background and Experience

- Could you please describe your academic background and your experience with qualitative data 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/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 model?
- 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?
- Example of a suggestion that does not make sense?
- Why do you think the sentence was suggested?

Text Tab

- What did you think of the way the keywords are visualised 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 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 model 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 with different stages/projects (top down approach vs bottom up approach)?

- 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?

Appendix B

Chapter 4 Supplementary Material

B.1 Manual Qualitative Data Analysis Themes and Codes

Table B.1: Themes (header) and codes identified in the manual thematic analysis performed on the interview transcripts in the user study.

Design Choices	Data Review	Using TACA/ML in QDA	Perception of the Model	User-Model Tensions	Interacting with the Model
Confusion tables	Commonalities	Automated coding	Accuracy interpretation	Ambiguity	Change from one theme to another
Confusion tables limitations	Correct	Comparisons with other tools	Accurate suggestions	Class imbalance	Drag and drop interaction
Confusion tables terminology	Data exploration	Easier	Confusion between keywords as pointers vs semantics	Confusion tables contradictions	Keywords as handles
Confusion tables tooltip	Explorative approach	Efficiency of using ML	Confusion tables agreements	Data quantity	Losing information
Different colors	Human fallibility	Efficient	Confusion tables strategy	Different performance on different parts of the data	Re-classification improvement
Frequency	Information extraction	Experience in qualitative analysis	Data convergence	Expectations	Re-classification limitations
Frequency counters	Keywords context	Faster	External source	Inaccuracy interpretation	Re-classification strategy
General limitations	Keywords to explore results	Human-tool distinction	Higher level of AI	Inaccurate suggestions	
Improvements	Manual coding	Iterative process	How the tool works	Justification of model inaccuracy	
Keywords	Manual coding process	Limited value of ML metrics	Improvement interpretation	Own lack of clarity	
Keywords tab	Own-coding feedback	ML influence on analysis	Interpretation	Trust	
Keywords limitations	Perception of things being objectively correct	Participant background	Keywords vs sentence embeddings	Trust in the model	
Keywords sorting	Reflexivity	Purpose	Lack of experience in ML	Trust in manual coding	
Keywords tables	Research question	Researcher vs tool distinction	Mental model	User own fault	
Keywords tooltip	Results summarization	Team collaboration	Missed sentences	User-model comparison	
Limitations of other tools	Semantic relations	Theme identification	Model evaluation		
Meaningless keywords	Semantics	Theme adjustments	Objective		
Performance limitations	Strategic approach	Useful	Perceived limitations of the model		
Setup	Subjectivity	Use case scenario	Personalized model		
Sources of confusion	Supportive role	User in control	Re-classification evaluation		
Text tab		Tedious	Transparency		
			Understanding		

Appendix C

Chapter 5 Supplementary Material

C.1 User Study Participant Information Sheet

Participant Information

Project Title: Interacting with a Text Classifier

We would like to invite you to participate in this research project directed by researchers at UCL. You should only participate if you want to; choosing not to take part will not disadvantage you in any way. Before you decide whether you want to take part, it is important for you to read the following information carefully and discuss it with others if you wish.

Study Details and Compensation

This study is part of a research project aiming to examine how users make sense of a text classifier. You will be exploring a Machine Learning (ML) system that has been trained on text (newspaper restaurant reviews). The text will not contain any offensive, personal, sexual or distasteful material.

If you agree to participate, you will be asked to complete a series of computer-based tasks. It is expected that the study will take no longer than **40 minutes**.

At the end of the activity, as compensation for your time, you will receive **£6**.

There are no particular risks associated with your participation other than those associated with the use of standard computer equipment.

Data and Information

All data will be handled according to the GDPR. **Any information** that is obtained in connection with this study and that can be identified with you will **remain confidential** and will be disclosed only with your permission or as required by law. Only UCL researchers working with Dr. Enrico Costanza will have access to data that can be identified with you.

Data in an aggregated or anonymous form (i.e. not revealing your identity), may instead be made publicly available through scientific publication, or otherwise shared with other researchers, as requested also by our funders (the UK Research Council).

The legal basis used to process your personal data will be the performance of a task in the public interest. The data controller for this project is University College London (UCL). The Data Protection Officer can be contacted at data-protection@ucl.ac.uk

Concerns or Complaints.

Should you have any concern or complaint, you can contact us at any point via email (federico.milana.18@ucl.ac.uk or e.costanza@ucl.ac.uk). For ethics queries, you can use the following contact (uclic-ethics@ucl.ac.uk). If you are concerned about how your personal data is being processed, you can contact the UCL Data Protection Office at data-protection@ucl.ac.uk. If you remain unsatisfied, you may wish to contact the Information Commissioner's Office (ICO). Contact details, and details of data subject rights, are available on the ICO website at <https://ico.org.uk/for-organisations/data-protection-reform/overview-of-the-gdpr/individuals-rights/>

Thank you for reading this information sheet and for considering taking part in this research study.

C.2 User Study Informed Consent Form

*Informed Consent Form

Project Title: Interacting with a Text Classifier.

Project ID No: UCLIC/1617/017/Staff Costanza/Nowacka/Yang

This study has been approved by the UCL Interaction Centre (UCLIC) Research Department's Ethics Committee.

Contact Details of Investigators

Principal investigator:

Dr. Enrico Costanza

UCL Gower Street

London WC1E 6BT

United Kingdom

+44 (0)20 7679 718

email: e.costanza@ucl.ac.uk

Co-investigator:

Federico Milana

UCL Gower Street

London WC1E 6BT

United Kingdom

email: federico.milana.18@ucl.ac.uk

Co-investigator:

Mirco Musolesi

UCL Gower Street

London WC1E 6BT

United Kingdom

email: m.musolesi@ucl.ac.uk

Co-investigator:

Amid Ayobi

UCL Gower Street

London WC1E 6BT

United Kingdom

email: amid.ayobi@ucl.ac.uk

Contact Details of the UCL Data Protection Officer

Alexandra Potts, data-protection@ucl.ac.uk

Participant's Statement

If you have any questions arising from the Information Sheet or explanation already given to you, please ask the researcher before you decide whether to join in.

I, the participant, confirm I understand that by agreeing to the items below I am consenting to participating in this study. I understand that it will be assumed that not checking the box mean that I DO NOT consent to that part of the study. I understand that by not giving consent for any one element I may be deemed ineligible for the study.

1. I confirm that I have read and understood the Information Sheet for the above study. I have had an opportunity to consider the information and what will be expected of me. I have also had the opportunity to ask questions which have been answered to my satisfaction and would like to take part in the survey study.
2. I consent to participate in the study. I understand that my personal information, in the form of this consent form, will be used for the purposes explained to me. I understand that according to data protection legislation, 'public task' will be the lawful basis for processing.
3. I understand that all personal information will remain confidential and that all efforts will be made to ensure I cannot be identified.
4. I understand that my data gathered in this study will be stored anonymously and securely. It will not be possible to identify me in any publications.
5. I understand that my information may be subject to review by responsible individuals from the University for monitoring and audit purposes.
6. I understand the potential risks of participating and the support that will be available to me should I become distressed during the course of the research.
7. I understand the direct/indirect benefits of participating.
8. I agree that my anonymized research data may be used by others for future research. (No one will be able to identify you when this data is shared).
9. I am aware of who I should contact if I wish to lodge a complaint
10. I agree for the data specified in the information sheet to be collected and stored on a secure server, and for an aggregate or anonymous version of this data to be made publicly available.

If you do not wish to participate in this study, please close this survey and return your submission on Prolific by selecting the 'Stop without completing' button.

I consent and agree to the terms (items 1-10) above

C.3 User Study Instructions (Heat Maps)

Hello, and thank you very much for participating in this study. Please read the following instructions carefully.

One of the successful applications of machine learning (ML) is text classification. It can be used to assign **labels** to words, sentences, or documents to make them easier to search, summarise or otherwise process by people. To achieve this, ML systems need to be **trained** on a large amount of text which is manually labelled. The text used for training is called the **training set**.

For this study, we trained an ML system to classify sentences from newspaper restaurant reviews into one or more of the following 4 topics, or **labels**:

1. food and drinks
2. place
3. people
4. opinions

So, if a sentence is about any of these topics, the system should assign it the corresponding label.

Please note that ML systems are generally not 100% accurate. They may work well on some sentences but make mistakes on others. The manual labelling used to train the ML system may also contain mistakes (e.g., we might have missed some labels for some sentences) and some sentences might be ambiguous, so different people might manually label them differently.

To measure its performance, an ML system is normally tested on sentences that were manually labelled, so we can check if the labels assigned by the ML system match those assigned manually.

There are typically four outcome measures:

1. A sentence was manually labelled as X (e.g., food and drinks) and the system labelled it as X – this is a **True Positive**.
2. A sentence was NOT manually labelled as X (e.g. food and drinks), and the system did NOT label it as X – this is a **True Negative**.
3. A sentence was labelled as X (e.g., food and drinks), but the system did NOT label it as X – this is a **False Negative**.
4. A sentence was NOT labelled as X (e.g., food and drinks), but the system labelled it as X – this is a **False Positive**.

Looking at examples for each of the outcomes can reveal how the system works. In this study, we ask you to look at a few examples and estimate how the system will perform.

Your main task in this study is to estimate whether the ML system will assign a given label to a given sentence. This task will be repeated 7 times for the label "opinions", and 7 times for the label "place". To help you with this, for each sentence you are asked to guess you will also be shown how the ML system performed on 12 sample sentences:

- 6 sentences that are **True Positive** examples
- 3 sentences that are **False Positive** examples
- 3 sentences that are **False Negative** examples

Afterwards, you will be asked about the topics or specific words that you think influence the system towards or against the label.

Finally, you will be given an additional sentence. You will be asked whether **you** would assign the label, and whether **the system** would assign the label.

For each True Positive, False Negative, and False Positive sentence, you will also be shown a **heat map**. Heat maps use different colours to show how each individual word contributes towards or against the label.

For example, here is a heat map for a sentence from a different data set. The sentence was manually labelled as "social", and the system has assigned the label "social" (True Positive):

"I think so, except for one time because I wasn't home, I was away in Paris and my flatmate was not here that weekend, so I think that you guys were trying to contact us, but I wasn't even seeing the messages because I was overseas."

Words highlighted in **red** contribute **towards** the label "social".

Words highlighted in **blue** contribute **against** the label "social".

Words that are not highlighted do not contribute towards or against the label "social".

Note that the shade of the colour indicates the strength of the word contribution: The **redder** the background and underline colour of the word, the more it contributes **towards** the label.

The **bluer** the background and underline colour of the word, the more it contributes **against** the label.

According to the heat map above, the words "contact" and "messages" have influenced the system to assign the label "social" to the sentence the most.

Conversely, the words "think", "time", and "trying" have influenced the system to NOT assign the label "social" the most (the label was still assigned).

C.4 User Study Instructions (Summaries)

Hello, and thank you very much for participating in this study. Please read the following instructions carefully.

One of the successful applications of machine learning (ML) is text classification. It can be used to assign **labels** to words, sentences, or documents to make them easier to search, summarise or otherwise process by people. To achieve this, ML systems need to be **trained** on a large amount of text which is manually labelled. The text used for training is called the **training set**.

For this study, we trained an ML system to classify sentences from newspaper restaurant reviews into one or more of the following **4 labels**:

1. food and drinks
2. place
3. people
4. opinions

So, if a sentence is about any of these topics, the system should assign it the corresponding label.

Please note that ML systems are generally not 100% accurate. They may work well on some sentences but make mistakes on others. The manual labelling used to train the ML system may also contain mistakes (e.g., we might have missed some labels for some sentences) and some sentences might be ambiguous, so different people might manually label them differently.

To measure its performance, an ML system is normally tested on sentences that were manually labelled, so we can check if the labels assigned by the ML system match those assigned manually.

There are typically four outcome measures:

1. A sentence was manually labelled as X (e.g., food and drinks) and the system labelled it as X – this is a **True Positive**.
2. A sentence was NOT manually labelled as X (e.g. food and drinks), and the system did NOT label it as X – this is a **True Negative**.
3. A sentence was labelled as X (e.g., food and drinks), but the system did NOT label it as X – this is a **False Negative**.
4. A sentence was NOT labelled as X (e.g., food and drinks), but the system labelled it as X – this is a **False Positive**.

Looking at examples for each of the outcomes can reveal how the system works. However, you might have to look at many examples to understand this. Instead, in this study, we ask you to read summaries of many examples and estimate how the system will perform.

Your main task in this study is to estimate whether the ML system will assign a given label to a given sentence. The task will be repeated 7 times for the label "opinions", and 7 times for the label "place".

To help you with this, we will show you summaries of words from sentences labelled by **the system**.

The summaries include the topics and words that support the classification, the topics and words that are neutral towards the classification, and those that go against the classification. These are according to **the system**, and what it has learned from the manually labelled training set.

Here is an **example** of a summary of words from sentences for the label "people":

Summary of the System Classification: "People"

Supporting Classification

The words that support the classification generally include names, positive actions, and terms related to organization or structure. For example, "allowed," "Stephen," "strikes," and "nicknamed." These words often imply permission, recognition, and activity which might be associated with positive reinforcement or facilitation within a structured system.

Neutral Towards Classification

The words that neither support nor oppose the classification include everyday objects, food items, and descriptive terms. Examples are "crystals," "chillies," "comforts," and "gooseberry." These words are mostly descriptive and relate to neutral elements that are part of daily life and culinary contexts, suggesting a middle ground that does not heavily influence the classification in either direction.

Against Classification

Words that are against the classification often carry connotations of imperfection, conflict, or negativity. Examples include "tragically," "mislaid," "ducting," and "smirk." These words tend to imply issues, errors, or negative judgments, which could suggest opposition to a positive or structured classification system.

Unique Characteristics

- **Supporting Words:** High frequency of proper nouns and action verbs which contribute to a dynamic and authoritative theme.
- **Neutral Words:** A diverse mix of everyday terms and descriptive adjectives indicating neutrality and ubiquity in everyday contexts.
- **Against Words:** Presence of negative connotations, verbs, and adjectives that imply problems or undesirable states.

In summary, the supporting words revolve around recognition, action, and structure. The neutral words are mundane and descriptive, often related to everyday life and culinary terms. The words against the classification tend to indicate problems, errors, and negative judgments.

C.5 LLM-Generated Summaries

C.5.1 Class: Opinions

Summary of the System Classification

Words Supporting the Classification

The words that support the classification are predominantly positive and emotionally charged. They include terms like "Love," "cares," "cheerful," "generous," and "joyous." These words convey sentiments of affection, kindness, and happiness. Additionally, some words such as "privilege," "rewarding," and "perfectly" suggest positive outcomes or states of being. A unique characteristic of this list is the presence of words that denote inclusivity and encouragement, such as "encourages" and "proud." The overarching theme for these words is positivity and emotional well-being.

Words Against the Classification

The words against the classification have a geographical and object-oriented theme. Examples include "lavender," "Walthamstow," "gnocchi," "Israel," "Persian," and "Cumberland." These terms are more specific and less emotionally charged, often referring to places, objects, or specific names. The unique characteristic here is the specificity and neutrality of the terms, which do not carry strong positive or negative connotations in an emotional sense but are rather factual or descriptive in nature.

Neutral Words

The neutral words, those that neither strongly support nor are against the classification, include terms like "associated," "aware," "technique," "recommendation," and "transparent." These words are largely descriptive and informational, focusing on relationships, states of being, and general descriptors. They lack the emotional charge found in the supporting words and the specificity of the against words. A unique characteristic of this list is its focus on general descriptions and states, indicating a lack of strong sentiment or bias.

Summary

In summary, the words supporting the classification are characterized by positivity and emotional well-being. The words against the classification are specific and descriptive, often related to geography and objects. The neutral words are general descriptors that provide information without strong emotional or specific connotations. This analysis highlights how different categories of words contribute to the overall sentiment and specificity within the classification.

C.5.2 Class: Place

Summary of the System Classification

Words Supporting Classification

The words that support the classification, such as "City," "seaside," "interior," "hotel," and "Greece," predominantly relate to locations, travel, and hospitality.

These words evoke images of places and experiences tied to geographical locations or tourist destinations. For example, "City," "seaside," and "hotel" suggest urban areas, coastal regions, and accommodations respectively. This indicates a theme centered around travel, tourism, and possibly the hospitality industry.

Words Against Classification

The words that are against the classification, such as "midst," "word," "undercooked," "vivid," and "succeed," are more abstract and not location-specific. They include terms related to quality or state, actions, and descriptions. Words like "undercooked" and "stupid" are negative descriptors, while "vivid" and "succeed" are more neutral or positive but still abstract. This suggests a theme of general descriptors and qualitative judgments that are not tied to specific locations or entities.

Neutral Words

The words that neither support nor are against the classification, such as "choral," "kitchen," "torched," "transparent," and "intrepid," are a mix of various categories. Some relate to specific objects or places ("kitchen," "steakhouse"), some describe states or qualities ("transparent," "modern"), and others describe actions or characteristics ("torched," "intrepid"). This variety indicates a lack of a specific common theme, reflecting a broad, generalized category without strong ties to either supporting or against classification themes.

Unique Characteristics

A unique characteristic that stands out is the strong presence of geographical and travel-related terms in the supporting classification group. This specificity contrasts sharply with the abstract and descriptive nature of the words in the against classification group. The neutral group's diversity highlights its role as a catch-all category, including terms from various unrelated domains.

C.5.3 Prompt Used

The following is a list of words ordered by their respective numerical weight in descending order. Words with higher weights support the classification. Words with lower weights are against the classification. The higher the weight, the more the word supports the classification. These words are at the top of the list. The lower the weight, the more the word is against the classification. These words are at the bottom of the list. The closer the weight is to zero, the more the word neither supports or is against the classification. These words are in the middle of the list. From all the words, extract a common theme for the words that support the classification, a common theme for the words that are against the classification, and a common theme for the words that neither support or are against the classification.

Highlight any unique characteristics that stand out. Write about 250 words.

C.6 Example User Study Page (LIME)

Label: Opinions 1/7

The label "opinions" includes the following topics:

opinions
personal thoughts
negative
positive
reviews
awards

True Positives

Examples of sentences that were manually labelled as "opinions", and the system also labelled as "opinions" (red words contribute towards "opinions", blue words against):

"She **worried**, when I invited her to join me, that it wouldn't be as **good** as other
Korean restaurants I've tried."

"I don't think **either** of those will have **provided**."

"It's probably the **best** £5 you can spend on lunch anywhere in Leeds **right now**."

"Clearly, they **need** the **staff**"

"It's **true** that dinner didn't get off to a **great** start."

"They will **think** I have made it **harder** for them to get a **table**."

False Negatives

Examples of sentences that were manually labelled as "opinions", but the system did **NOT** label as "opinions" (red words contribute **towards** "opinions", blue words **against**):

"why, **bar** a **small** mention in Olive magazine, had I found **nothing** else about it in
the national press since it opened in 2018?"

"At those prices I only **wanted** lunch, not a **parlour** game."

"That **said**, I do **give** **thanks** for the **modern**, **industrial-scale** **extractor** system that
dangles down over **every** **table** from the **high** **ceiling** here at the back"

False Positives

Examples of sentences that were **NOT** manually labelled as "opinions" but the system labelled as "opinions" (red words contribute **towards** "opinions", blue words **against**):

"I **would** have had to **spend** the following **paragraphs** devising some **desperate**
excuse for why I **ordered** what **would** have been by **far** the most **expensive** dish
available"

"There are **still** a lot of **businesses** taking **advantage** of some **beneficial** post-
Covid **leases**, but if it also **means** the likes of Hampton and Lovell try their hands
here instead, on the **site** of what was a Café Rouge I **never** **visited**, then I'm all for
it."

"If you **fancy**, they will **tell** you **exactly** what you're being **served** as each **plate**
arrives, but there's also a **card** on the **table** inviting you to 'challenge your **foodie**
senses and **write** what you can taste in each **course'** "

Questions

Based on the examples shown to you:

*Are there any topics or words that influence the system **towards** the label "opinions"? If so, please name 1-3 topics or words:

*Are there any topics or words that are **not** considered by the system when labelling sentences as "opinions"? (i.e. topics or words that are neither for or against the classification) If so, please name 1-3 topics or words:

"I ordered it because it's my job to do so and, as ever, so you wouldn't have to"

*Would **you** assign the label "opinions" to this sentence?

Yes

No

*How confident are you in your answer?

Extremely unconfident

Slightly unconfident

Slightly confident

Extremely confident

*Do you think **the system** would assign the label "opinions" to this sentence?

Yes

No

*How confident are you in your answer?

Extremely unconfident

Slightly unconfident

Slightly confident

Extremely confident

C.7 Prediction Accuracy Across Different Interpretability Techniques

Table C.1: Comparison of prediction accuracy based on different interpretability techniques for all tasks.

Task	Class Label	Category	No-Exp.	LIME	SHAP	Summaries	Chi-Square	p-value
“I ordered it because it’s my job to do so and, as ever, so you wouldn’t have to”	Opinions	TP	0.59	0.38	0.53	0.50	3.25 (3,128)	0.36
“It turns out their style is to be extremely mean with the garlic butter, which is the whole point of eating snails”	Opinions	TP	0.72	0.72	0.81	0.72	1.10 (3, 128)	0.78
“The soup element is missing in action”	Opinions	TP	0.63	0.66	0.63	0.47	2.85 (3,128)	0.42
“Wine lists full of triple digits land first with the cocktail list, when what you really want is the menu”	Opinions	FN	0.41	0.59	0.34	0.47	4.41 (3, 128)	0.22
“The £12 salade niçoise is a pleasingly dense, chopped affair, with a pretty arrangement of anchovies and tomatoes, topped by half a boiled egg, its yolk at a perfect state of jamminess”	Opinions	FN	0.25	0.22	0.19	0.09	2.87 (3, 128)	0.41
“Upstairs, past the spray-painted words ‘liberté, égalité, fraternité’, is a dining room seating another mighty 20.”	Opinions	FP	0.31	0.31	0.19	0.19	2.67 (3, 128)	0.45
“It was a place where the tables were apparently so difficult to nab even the waiting list had a waiting list.”	Opinions	FP	0.66	0.66	0.59	0.50	2.18 (3, 128)	0.54
“Fancy Mayfair restaurants are full of older people wearing young people’s shirts”	Place	TP	0.53	0.78	0.72	0.78	6.44 (3, 128)	0.09
“drink something sweetly familiar from the short wine list that hasn’t heard of anywhere outside France”	Place	TP	0.75	0.75	0.81	0.63	3.04 (3, 128)	0.39
“But the people he was feeding at the Chiltern Firehouse were only there to see and be obscene.”	Place	TP	0.75	0.72	0.63	0.88	5.35 (3, 128)	0.15
“The £26 salade Niçoise is lovely in a quiet, understated way.”	Place	FN	0.53	0.69	0.69	0.72	3.05 (3, 128)	0.38
“The trout dish is £24.”	Place	FN	0.22	0.44	0.50	0.78	20.65 (3, 128)	< 0.001
“they told me that seven people had been comped that lunchtime and eight had not, mostly because they declined the offer”	Place	FP	0.28	0.28	0.34	0.13	4.37 (3, 128)	0.22
“Mendes is also the executive chef”	Place	FP	0.34	0.13	0.16	0.06	9.88 (3, 128)	0.02

C.8 Pilot Studies Results

C.8.1 Pilot Study 1

C.8.1.1 Prediction Accuracy on Sampled Tasks

Table C.2: Comparison of prediction accuracy based on different interpretability techniques for all tasks in the first pilot study.

Task	Class Label	Classified	No-Exp.	LIME	SHAP	Summaries
“Ours comes crusted with cinnamon-boosted sugar with a bowl of soft serve ice-cream, caramelised popcorn and a little fruit.”	Food & Drinks	Yes	1.00	1.00	1.00	1.00
“There’s a truly terrible dish of undercooked aubergine, with a bland buttermilk dressing that tastes of very little.”	Food & Drinks	Yes	0.88	0.94	0.94	1.00
“There are six of us at the table and nine dishes, so let’s find out”	Food & Drinks	No	0.81	0.75	0.81	0.94
“Finally, in April 2021, his restaurant reopened”	Food & Drinks	No	0.94	0.69	0.75	0.81
“Obviously, the pandemic intervened, during which Diagne cooked in a local church for vulnerable people in his community.”	People	Yes	0.81	0.81	0.88	0.88
“Eventually he found sanctuary in a church and then got into a hostel.”	People	Yes	0.31	0.31	0.38	0.44
“What matters is that Coventry has this rough wood panel-clad space offering a very good time to anyone up for the joys of cooking their own lunch.”	People	No	0.56	0.75	0.69	0.75
“It comes on a deep green herby emulsion, which in turn is on a piece of monogrammed paper.”	People	No	1.00	0.94	0.88	0.81

C.8.1.2 Distribution of Correct Answers by Condition

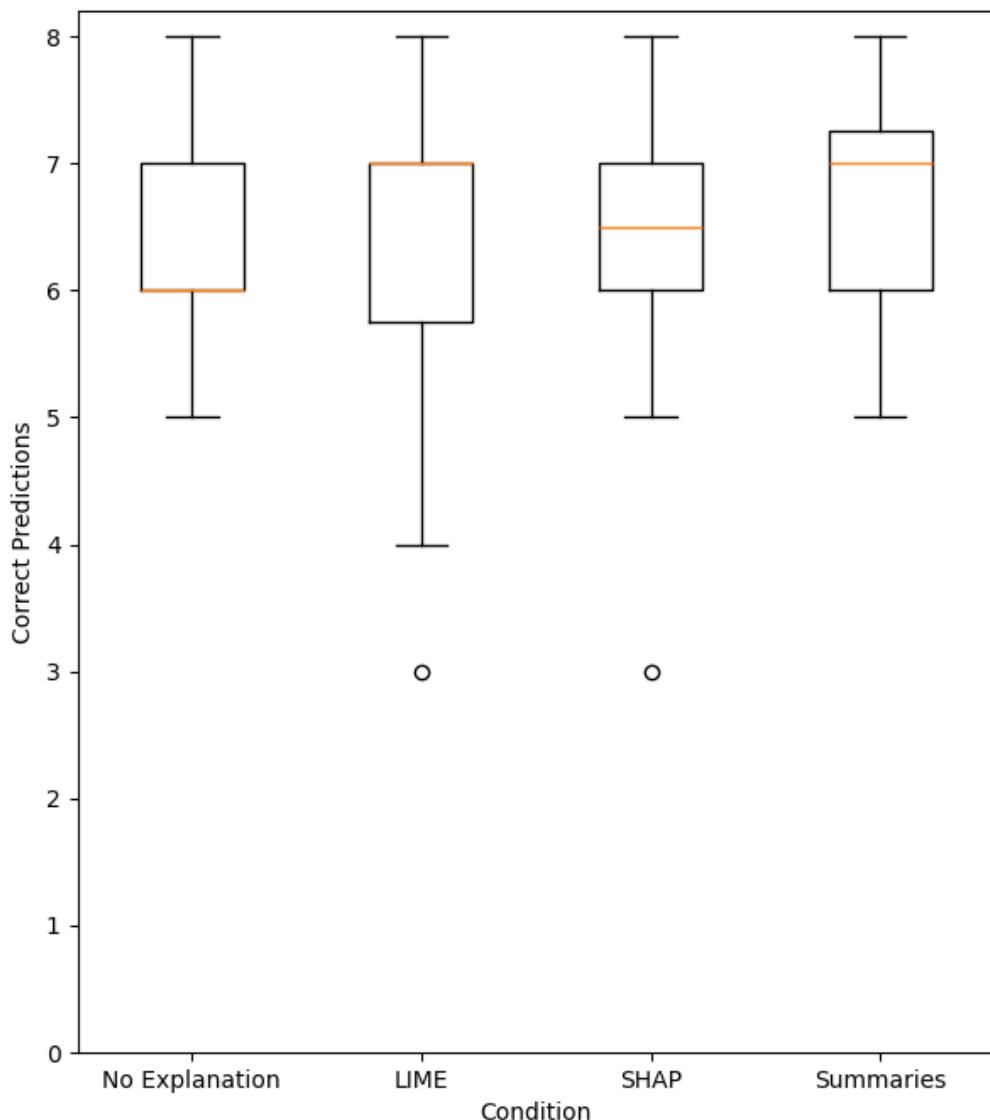


Figure C.1: Distributions of total correct predictions to all questions by condition in the first pilot study.

C.8.1.3 ANOVA Results for Correct Predictions by Condition

Table C.3: ANOVA results for the effect of condition on the number of correct predictions in the first pilot study. Note: Type III Sum of Squares.

Cases	Sum of Squares	df	Mean Square	F	p
Condition	1.375	3	0.458	0.345	0.793
Residuals	79.625	60	1.327		

C.8.2 Pilot Study 2

C.8.2.1 Prediction Accuracy on Sampled Tasks

Table C.4: Comparison of prediction accuracy based on different interpretability techniques for all tasks in the second pilot study.

Task	Class Label	Category	No-Exp.	LIME	SHAP	Summaries
“The womb of a red-walled dining room feels like a happy place, the rhythm of the chatter marked out by the clatter of knife and fork on plate.”	Food & Drinks	FP	0.56	0.56	0.50	0.25
“Now I’ve discovered I love it, which is a relief.”	Food & Drinks	FP	0.06	0.06	0.06	0.13
“Main courses justify their £20-plus price tags by both execution and volume”	Food & Drinks	FN	0.56	0.31	0.25	0.63
“There’s a list of steaks to be cooked on the clanking Argentine asador grill, on display behind the huge plateglass window at the back”	Food & Drinks	FN	0.00	0.06	0.00	0.00
“But he’s clearly also extremely good at the laidback and enfolding.”	People	FP	0.44	0.38	0.50	0.63
“The skewered folds of beef served there had been dry and flavourless, unless introduced to the pile of spices sitting far to the side of a plate, like an overly exuberant party guest kept away from everyone else for fear they’ll disgrace themselves.”	People	FP	0.31	0.13	0.56	0.44
“It feeds and it cares.”	People	FN	0.19	0.44	0.38	0.13
“Ah, here’s our waiter who is charming and efficient, and cursed with having to tell us, with great solemnity, that we are about to be taken on ’A Mediterranean Culinary Odyssey’”	People	FN	0.31	0.13	0.19	0.19

C.8.2.2 Distribution of Correct Answers by Condition

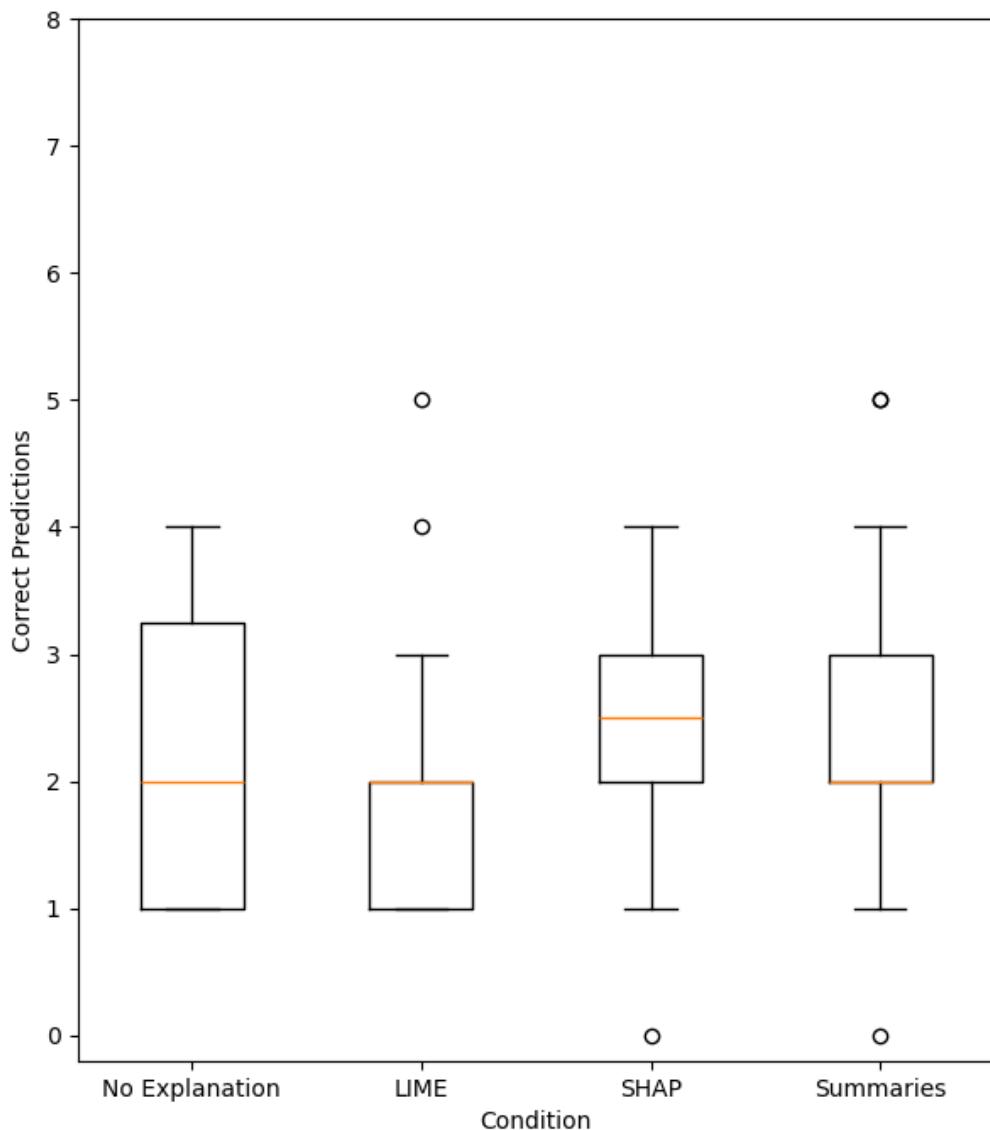


Figure C.2: Distributions of total correct predictions to all questions by condition in the second pilot study.

C.8.2.3 ANOVA Results for Correct Predictions by Condition

Table C.5: ANOVA results for the effect of condition on correct predictions in the second pilot study. Note: Type III Sum of Squares.

Cases	Sum of Squares	df	Mean Square	F	p
Condition	1.688	3	0.563	0.426	0.735
Residuals	79.250	60	1.321		

Bibliography

- Abdul, Ashraf et al. (2018). “Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI ’18. Montreal QC, Canada: Association for Computing Machinery, pp. 1–18. ISBN: 9781450356206. DOI: 10.1145/3173574.3174156.
- Adadi, Amina and Mohammed Berrada (2018). “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)”. In: *IEEE Access* 6. Conference Name: IEEE Access, pp. 52138–52160. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2018.2870052.
- Adams, Tony E (2008). “A review of narrative ethics”. In: *Qualitative inquiry* 14.2. Publisher: Sage Publications Sage CA: Los Angeles, CA, pp. 175–194.
- Adams, Tony E., Carolyn Ellis, and Stacy Holman Jones (2017). “Autoethnography”. en. In: *The International Encyclopedia of Communication Research Methods*. John Wiley and Sons, Ltd, pp. 1–11. ISBN: 978-1-118-90173-1. DOI: 10.1002/9781118901731.iecrm0011.
- Aechtner, Jonathan et al. (July 2022). “Comparing User Perception of Explanations Developed with XAI Methods”. In: *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–7. DOI: 10.1109/FUZZ-IEEE55066.2022.9882743.
- Alalouch, Chaham (Sept. 2021). “Cognitive Styles, Gender, and Student Academic Performance in Engineering Education”. en. In: *Education Sciences* 11.9. Number: 9 Publisher: Multidisciplinary Digital Publishing Institute, p. 502. ISSN: 2227-7102. DOI: 10.3390/educsci11090502.

- Alqaraawi, Ahmed et al. (Mar. 2020). “Evaluating Saliency Map Explanations for Convolutional Neural Networks: A User Study”. In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. IUI ’20. New York, NY, USA: Association for Computing Machinery, pp. 275–285. ISBN: 978-1-4503-7118-6. DOI: 10.1145/3377325.3377519.
- Alvarado, Oscar 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. New York, NY, USA: Association for Computing Machinery, pp. 1–12. ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3173860. URL: <https://doi.org/10.1145/3173574.3173860>.
- Amershi, Saleema (2011). “Designing for effective end-user interaction with machine learning”. In: *Proceedings of the 24th Annual ACM Symposium Adjunct on User Interface Software and Technology*. UIST ’11 Adjunct. Santa Barbara, California, USA: Association for Computing Machinery, pp. 47–50. ISBN: 9781450310147. DOI: 10.1145/2046396.2046416.
- Amershi, Saleema, Maya Cakmak, et al. (Dec. 2014). “Power to the People: The Role of Humans in Interactive Machine Learning”. en. In: *AI Magazine* 35.4. Number: 4, pp. 105–120. ISSN: 2371-9621. DOI: 10.1609/aimag.v35i4.2513.
- Amershi, Saleema, Max Chickering, et al. (Apr. 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. New York, NY, USA: Association for Computing Machinery, pp. 337–346. ISBN: 978-1-4503-3145-6. DOI: 10.1145/2702123.2702509.
- Amershi, Saleema, James Fogarty, et al. (Oct. 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. New York, NY, USA: Association for Computing Machinery, pp. 247–256. ISBN: 978-1-60558-745-5. DOI: 10.1145/1622176.1622222.

- Amershi, Saleema, Dan Weld, et al. (2019). “Guidelines for Human-AI Interaction”. In: CHI ’19. Glasgow, Scotland Uk: Association for Computing Machinery, pp. 1–13. ISBN: 9781450359702. DOI: 10.1145/3290605.3300233.
- Andrews, Christopher, Alex Endert, and Chris North (Apr. 2010). “Space to think: large high-resolution displays for sensemaking”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’10. New York, NY, USA: Association for Computing Machinery, pp. 55–64. ISBN: 978-1-60558-929-9. DOI: 10.1145/1753326.1753336.
- Aoki, Paul M (2007). “Back stage on the front lines: perspectives and performance in the combat information center”. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 717–726.
- Arendt, Dustin et al. (Mar. 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. New York, NY, USA: Association for Computing Machinery, pp. 591–602. ISBN: 978-1-4503-6272-6. DOI: 10.1145/3301275.3302280.
- Arvidsson, Matilda and Gregor Noll (2023). “Decision Making in Asylum Law and Machine Learning: Autoethnographic Lessons Learned on Data Wrangling and Human Discretion”. In: *Nordic Journal of International Law* 92.1, pp. 56–92.
- Ashley, Kevin D (2017). *Artificial intelligence and legal analytics: new tools for law practice in the digital age*. Cambridge University Press.
- El-Assady, Mennatallah et al. (2017). “NEREx: Named-Entity Relationship Exploration in Multi-Party Conversations”. en. In: *Computer Graphics Forum* 36.3, pp. 213–225. ISSN: 1467-8659. DOI: 10.1111/cgf.13181.
- Atanasova, Pepa et al. (Sept. 2020). *A Diagnostic Study of Explainability Techniques for Text Classification*. arXiv:2009.13295 [cs]. DOI: 10.48550/arXiv.2009.13295.
- Ayobi, Amid et al. (2021). “Machine Learning Explanations as Boundary Objects: How AI Researchers Explain and Non-Experts Perceive Machine Learning”.

- en. In: *2021 Joint ACM Conference on Intelligent User Interfaces Workshops, ACMIUI-WS 2021*. CEUR Workshop Proceedings.
- Baharudin, Baharum et al. (Feb. 2010). “A Review of Machine Learning Algorithms for Text-Documents Classification”. In: *Journal of Advances in Information Technology* 1. DOI: 10.4304/jait.1.1.4-20.
- Bang, Tove Grimstad et al. (July 2024). “A Retrospective Autoethnography Documenting Dance Learning Through Data Physicalisations”. In: *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. DIS ’24. New York, NY, USA: Association for Computing Machinery, pp. 2357–2373. ISBN: 9798400705830. DOI: 10.1145/3643834.3661607.
- Bansal, Gagan et al. (2021). “Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. Yokohama, Japan: Association for Computing Machinery. ISBN: 9781450380966. DOI: 10.1145/3411764.3445717.
- Bartleet, Brydie-Leigh (2021). “Artistic autoethnography: Exploring the interface between autoethnography and artistic research”. In: *Handbook of autoethnography*. Routledge, pp. 133–145.
- Beauxis-Aussalet, Emma and Lynda Hardman (2014). “Visualization of confusion matrix for non-expert users”. In: *IEEE Conference on Visual Analytics Science and Technology (VAST)-Poster Proceedings*. sn, pp. 1–2.
- Belle, Vaishak and Ioannis Papantonis (2021). “Principles and Practice of Explainable Machine Learning”. In: *Frontiers in Big Data* 4. ISSN: 2624-909X. DOI: 10.3389/fdata.2021.688969.
- Berg, Stuart et al. (2019). “Ilastik: interactive machine learning for (bio) image analysis”. In: *Nature methods* 16.12, pp. 1226–1232.
- Bergman, Eric and R Haitani (2000). “Designing the palm pilot”. In: *Information appliances and beyond*. Publisher: Morgan Kaufmann San Francisco, pp. 81–102.

- Bernard, Jürgen, Marco Hutter, et al. (Jan. 2018). “Comparing Visual-Interactive Labeling with Active Learning: An Experimental Study”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1. Conference Name: IEEE Transactions on Visualization and Computer Graphics, pp. 298–308. ISSN: 1941-0506. DOI: 10.1109/TVCG.2017.2744818.
- Bernard, Jürgen, Matthias Zeppelzauer, et al. (Sept. 2018). “VIAL: a unified process for visual interactive labeling”. en. In: *The Visual Computer* 34.9, pp. 1189–1207. ISSN: 1432-2315. DOI: 10.1007/s00371-018-1500-3.
- Bochner, Arthur P. (1984). “The Functions of Communication in Interpersonal Bonding”. In: *Handbook of Rhetorical and Communication Theory*. Ed. by C. C. Arnold. Boston: Allyn & Bacon, pp. 544–621.
- (1994). “Perspectives on Inquiry II: Theories and Stories”. In: *Handbook of Interpersonal Communication*. Ed. by M. L. Knapp and G. R. Miller. 2nd. Thousand Oaks, CA: Sage Publications, pp. 21–41.
- (2002). “Perspectives on Inquiry III: The Moral of Stories”. In: *Handbook of Interpersonal Communication*. Ed. by M. L. Knapp and J. A. Daly. 3rd. Thousand Oaks, CA: Sage Publications, pp. 73–101.
- Boukhelifa, Nadia, Anastasia Bezerianos, and Evelyne Lutton (2018). “Evaluation of Interactive Machine Learning Systems”. en. In: *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*. Ed. by Jianlong Zhou and Fang Chen. Human–Computer Interaction Series. Cham: Springer International Publishing, pp. 341–360. ISBN: 978-3-319-90403-0. DOI: 10.1007/978-3-319-90403-0_17.
- Bowker, Geoffrey and Susan Leigh Star (1999). “Sorting things out”. In: *Classification and its consequences* 4. Publisher: Citeseer.
- Braun, Virginia and Victoria Clarke (Jan. 2006). “Using thematic analysis in psychology”. In: *Qualitative Research in Psychology* 3.2, pp. 77–101. ISSN: 1478-0887. DOI: 10.1191/1478088706qp063oa.
- Braun, Virginia and Victoria Clarke (Aug. 2019). “Reflecting on reflexive thematic analysis”. In: *Qualitative Research in Sport, Exercise and Health* 11.4,

- pp. 589–597. ISSN: 2159-676X. DOI: 10 . 1080 / 2159676X . 2019 . 1628806.
- Brauner, Philipp et al. (2023). “What does the public think about artificial intelligence?—A criticality map to understand bias in the public perception of AI”. In: *Frontiers in Computer Science* 5. ISSN: 2624-9898. DOI: 10 . 3389 / fcomp . 2023 . 1113903.
- Buchenau, Marion and Jane Fulton Suri (Aug. 2000). “Experience prototyping”. In: *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques*. DIS '00. New York, NY, USA: Association for Computing Machinery, pp. 424–433. ISBN: 978-1-58113-219-9. DOI: 10 . 1145 / 347642 . 347802.
- Buçinca, Zana, Phoebe Lin, et al. (2020). “Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems”. In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. IUI '20. Cagliari, Italy: Association for Computing Machinery, pp. 454–464. ISBN: 9781450371186. DOI: 10 . 1145 / 3377325 . 3377498.
- Buçinca, Zana, Maja Barbara Malaya, and Krzysztof Z. Gajos (Apr. 2021). “To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making”. In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW1. DOI: 10 . 1145 / 3449287.
- Bulathwela, Sahan et al. (2024). “Artificial Intelligence Alone Will Not Democratisate Education: On Educational Inequality, Techno-Solutionism and Inclusive Tools”. In: *Sustainability* 16.2. ISSN: 2071-1050. DOI: 10 . 3390 / su16020781.
- Bussone, Adrian, Simone Stumpf, and Dympna O’Sullivan (2015). “The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems”. In: *2015 International Conference on Healthcare Informatics*, pp. 160–169. DOI: 10 . 1109 / ICHI . 2015 . 26.

- Byrne, David (June 2022). “A worked example of Braun and Clarke’s approach to reflexive thematic analysis”. en. In: *Quality and Quantity* 56.3, pp. 1391–1412. ISSN: 1573-7845. DOI: 10.1007/s11135-021-01182-y.
- Cai, Carrie J., Jonas Jongejan, and Jess Holbrook (Mar. 2019). “The Effects of Example-Based Explanations in a Machine Learning Interface”. In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI ’19. New York, NY, USA: Association for Computing Machinery, pp. 258–262. ISBN: 978-1-4503-6272-6. DOI: 10.1145/3301275.3302289.
- Caramiaux, Baptiste, Fabien Lotte, et al. (Apr. 2019). *AI in the media and creative industries*. Research Report. New European Media (NEM), pp. 1–35. URL: <https://inria.hal.science/hal-02125504>.
- Caramiaux, Baptiste and Atau Tanaka (2013). “Machine learning of musical gestures: Principles and review”. In: *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*. Graduate School of Culture Technology, KAIST, pp. 513–518.
- Carmichael, Josie et al. (Mar. 2024). “Diagnostic decisions of specialist optometrists exposed to ambiguous deep-learning outputs”. en. In: *Scientific Reports* 14.1. Publisher: Nature Publishing Group, p. 6775. ISSN: 2045-2322. DOI: 10.1038/s41598-024-55410-0.
- Caruana, R. et al. (1999). “Case-Based Explanation of Non-Case-Based Learning Methods”. In: *Proceedings of the AMIA Symposium*, pp. 212–215. ISSN: 1531-605X. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2232607/>.
- Cesarini, Mirko et al. (2024). “Explainable ai for text classification: Lessons from a comprehensive evaluation of post hoc methods”. In: *Cognitive Computation*, pp. 1–19.
- Charte, Francisco et al. (Nov. 2015). “MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation”. en. In: *Knowledge-Based Systems* 89, pp. 385–397. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2015.07.019.

- Chen, Nan-Chen et al. (June 2018). “Using Machine Learning to Support Qualitative Coding in Social Science: Shifting the Focus to Ambiguity”. In: *ACM Transactions on Interactive Intelligent Systems* 8.2, 9:1–9:20. ISSN: 2160-6455. DOI: 10.1145/3185515.
- Choi, May Y. and Christopher Ma (Aug. 2020). “Making a big impact with small datasets using machine-learning approaches”. English. In: *The Lancet Rheumatology* 2.8, e451–e452. DOI: 10.1016/S2665-9913(20)30217-4.
- Chromik, Michael et al. (Apr. 2021). “I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI”. In: *Proceedings of the 26th International Conference on Intelligent User Interfaces*. IUI ’21. New York, NY, USA: Association for Computing Machinery, pp. 307–317. ISBN: 978-1-4503-8017-1. DOI: 10.1145/3397481.3450644.
- Chuang, Jason et al. (May 2012). “Interpretation and trust: designing model-driven visualizations for text analysis”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’12. New York, NY, USA: Association for Computing Machinery, pp. 443–452. ISBN: 978-1-4503-1015-4. DOI: 10.1145/2207676.2207738.
- Collaris, Dennis, Leo M. Vink, and Jarke J. van Wijk (June 2018). *Instance-Level Explanations for Fraud Detection: A Case Study*. arXiv:1806.07129 [cs, stat]. DOI: 10.48550/arXiv.1806.07129.
- Collins, Christopher, Sheelagh Carpendale, and Gerald Penn (2009). “DocuBurst: Visualizing Document Content using Language Structure”. en. In: *Computer Graphics Forum* 28.3, pp. 1039–1046. ISSN: 1467-8659. DOI: 10.1111/j.1467-8659.2009.01439.x.
- Corbett, E. and N. Saul (Oct. 2018). “Interactive Machine Learning Heuristics in Learning from Users”. en. In: *IEEE Vis Workshop*.
- Covert, Ian, Scott M Lundberg, and Su-In Lee (2020). “Understanding Global Feature Contributions With Additive Importance Measures”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al.

- Vol. 33. Curran Associates, Inc., pp. 17212–17223. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/c7bf0b7c1a86d5eb3be2c722cf2cf746-Paper.pdf.
- Craven, Mark and Jude Shavlik (1995). “Extracting Tree-Structured Representations of Trained Networks”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Touretzky, M.C. Mozer, and M. Hasselmo. Vol. 8. MIT Press. URL: https://proceedings.neurips.cc/paper_files/paper/1995/file/45f31d16b1058d586fc3be7207b58053-Paper.pdf.
- Crowston, Kevin, Eileen Allen, and Robert Heckman (Nov. 2012). “Using natural language processing technology for qualitative data analysis”. In: *International Journal of Social Research Methodology* 15.6. Publisher: Routledge eprint: <https://doi.org/10.1080/13645579.2011.625764>, pp. 523–543. ISSN: 1364-5579. DOI: 10.1080/13645579.2011.625764.
- Crowston, Kevin, Xiaozhong Liu, and Eileen Allen (Nov. 2010). “Machine Learning and Rule-Based Automated Coding of Qualitative Data”. In: *Proceedings of the American Society for Information Science and Technology* 47, pp. 1–2. DOI: 10.1002/meet.14504701328.
- De Certeau, Michel and Steven F Rendall (2004). “from The Practice of Every-day Life (1984)”. In: *The City Cultures Reader* 3.2004. Publisher: Psychology Press, p. 266.
- Delgado-Panadero, Ángel et al. (Apr. 2022). “Implementing local-explainability in Gradient Boosting Trees: Feature Contribution”. In: *Information Sciences* 589, pp. 199–212. ISSN: 0020-0255. DOI: 10.1016/j.ins.2021.12.111.
- Desjardins, Audrey and Aubree Ball (June 2018). “Revealing Tensions in Autobiographical Design in HCI”. In: *Proceedings of the 2018 Designing Interactive Systems Conference*. DIS ’18. New York, NY, USA: Association for Computing Machinery, pp. 753–764. ISBN: 978-1-4503-5198-0. DOI: 10.1145/3196709.3196781.

- Desjardins, Audrey, Oscar Tomico, et al. (Dec. 2021). “Introduction to the Special Issue on First-Person Methods in HCI”. In: *ACM Trans. Comput.-Hum. Interact.* 28.6. ISSN: 1073-0516. DOI: 10.1145/3492342.
- Devlin, Jacob et al. (May 2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv:1810.04805 [cs]. DOI: 10.48550/arXiv.1810.04805.
- Dhurandhar, Amit et al. (2018). “Improving Simple Models with Confidence Profiles”. In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc. URL: <https://proceedings.neurips.cc/paper/2018/hash/972cdale62b72640cb7ac702714a115f-Abstract.html> (visited on 08/26/2024).
- Ding, Weiping, Mohamed Abdel-Basset, and Ahmed M. Hawash Hossam Ali (2022). “Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey”. In: *Information Sciences* 615, pp. 238–292. ISSN: 0020-0255. DOI: <https://doi.org/10.1016/j.ins.2022.10.013>.
- Doshi-Velez, Finale and Been Kim (Mar. 2017). *Towards A Rigorous Science of Interpretable Machine Learning*. arXiv:1702.08608 [cs, stat]. DOI: 10.48550/arXiv.1702.08608.
- Duan, Tony et al. (July 2020). “NGBoost: Natural Gradient Boosting for Probabilistic Prediction”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 2690–2700. URL: <https://proceedings.mlr.press/v119/duan20a.html>.
- Dudley, John J. and Per Ola Kristensson (June 2018). “A Review of User Interface Design for Interactive Machine Learning”. In: *ACM Transactions on Interactive Intelligent Systems* 8.2, 8:1–8:37. ISSN: 2160-6455. DOI: 10.1145/3185517.

- Ecclesia, Vittoria (2023). “ART AND RIGOUR: CREATING METHODOLOGIES FOR ARTISTIC RESEARCH IN MUSIC”. In: *Culture Crossroads* 22.1, pp. 12–21.
- Ehsan, Upol et al. (2022). “Human-Centered Explainable AI (HCXAI): Beyond Opening the Black-Box of AI”. In: *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI EA ’22. New Orleans, LA, USA: Association for Computing Machinery. ISBN: 9781450391566. DOI: 10.1145/3491101.3503727.
- Ellis, Carolyn (2004). *The ethnographic I: A methodological novel about autoethnography*. Rowman Altamira.
- Ellis, Carolyn, Tony E. Adams, and Arthur P. Bochner (2011). “Autoethnography: An Overview”. In: *Historical Social Research / Historische Sozialforschung* 36.4 (138), pp. 273–290. ISSN: 01726404. URL: <http://www.jstor.org/stable/23032294> (visited on 11/10/2024).
- Endert, Alex, Patrick Fiaux, and Chris North (May 2012). “Semantic interaction for visual text analytics”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’12. New York, NY, USA: Association for Computing Machinery, pp. 473–482. ISBN: 978-1-4503-1015-4. DOI: 10.1145/2207676.2207741.
- Fails, Jerry Alan and Dan R. Olsen (Jan. 2003). “Interactive machine learning”. In: *Proceedings of the 8th international conference on Intelligent user interfaces*. IUI ’03. New York, NY, USA: Association for Computing Machinery, pp. 39–45. ISBN: 978-1-58113-586-2. DOI: 10.1145/604045.604056.
- Faith, Jeremiah (2024). “Through the AI thinking space: An autoethnographic tale of unexpected insight”. In.
- Fassl, Matthias and Katharina Krombholz (Apr. 2023). “Why I Can’t Authenticate - Understanding the Low Adoption of Authentication Ceremonies with Autoethnography”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. New York, NY, USA: Association for

- Computing Machinery, pp. 1–15. ISBN: 978-1-4503-9421-5. DOI: 10.1145/3544548.3581508.
- Fisher, Walter R (1984). “Narration as a human communication paradigm: The case of public moral argument”. In: *Communications Monographs* 51.1. Publisher: Taylor and Francis, pp. 1–22.
- Flick, Uwe (2020). “Gütekriterien qualitativer Forschung”. In: *Handbuch Qualitative Forschung in der Psychologie: Band 2: Designs und Verfahren*. Publisher: Springer, pp. 247–263.
- Fogarty, James et al. (Apr. 2008). “CueFlik: interactive concept learning in image search”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’08. New York, NY, USA: Association for Computing Machinery, pp. 29–38. ISBN: 978-1-60558-011-1. DOI: 10.1145/1357054.1357061.
- Fontana, Joyce S (2004). “A methodology for critical science in nursing”. In: *Advances in nursing science* 27.2, pp. 93–101.
- Gaver, William and Frances Gaver (Apr. 2023). “Living with Light Touch: An Autoethnography of a Simple Communication Device in Long-Term Use”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. New York, NY, USA: Association for Computing Machinery, pp. 1–14. ISBN: 978-1-4503-9421-5. DOI: 10.1145/3544548.3580807.
- Gebreegziabher, Simret Araya et al. (Apr. 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. New York, NY, USA: Association for Computing Machinery, pp. 1–19. ISBN: 978-1-4503-9421-5. DOI: 10.1145/3544548.3581352.
- Gebru, Timnit et al. (2022). “Excerpt from Datasheets for Datasets”. In: *Ethics of Data and Analytics*. Num Pages: 9. Auerbach Publications. ISBN: 978-1-00-327829-0.
- Geva, Mor, Yoav Goldberg, and Jonathan Berant (Aug. 2019). *Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Lan-*

- guage Understanding Datasets. arXiv:1908.07898 [cs]. DOI: 10 . 48550 / arXiv.1908.07898.
- Gillies, Marco et al. (2016). “Human-Centred Machine Learning”. In: *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. CHI EA ’16. San Jose, California, USA: Association for Computing Machinery, pp. 3558–3565. ISBN: 9781450340823. DOI: 10 . 1145 / 2851581.2856492.
- Glikson, Ella and Anita Williams Woolley (2020). “Human trust in artificial intelligence: Review of empirical research”. In: *Academy of Management Annals* 14.2, pp. 627–660.
- Gorwa, Robert, Reuben Binns, and Christian Katzenbach (2020). “Algorithmic content moderation: Technical and political challenges in the automation of platform governance”. In: *Big Data & Society* 7.1, p. 2053951719897945.
- Grimmer, Justin and Brandon M. Stewart (2013). “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts”. en. In: *Political Analysis* 21.3, pp. 267–297. ISSN: 1047-1987, 1476-4989. DOI: 10 . 1093/pan/mps028.
- Groce, Alex et al. (Mar. 2014). “You Are the Only Possible Oracle: Effective Test Selection for End Users of Interactive Machine Learning Systems”. In: *IEEE Transactions on Software Engineering* 40.3, pp. 307–323. ISSN: 1939-3520. DOI: 10 . 1109/TSE . 2013 . 59.
- Guidotti, Riccardo et al. (Aug. 2018). “A Survey of Methods for Explaining Black Box Models”. In: *ACM Comput. Surv.* 51.5, 93:1–93:42. ISSN: 0360-0300. DOI: 10 . 1145 / 3236009.
- Gupta, Shivani and Atul Gupta (Jan. 2019). “Dealing with Noise Problem in Machine Learning Data-sets: A Systematic Review”. In: *Procedia Computer Science*. The Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia 161, pp. 466–474. ISSN: 1877-0509. DOI: 10 . 1016/j.procs.2019.11.146.

- Hartmann, Björn et al. (Apr. 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. New York, NY, USA: Association for Computing Machinery, pp. 145–154. ISBN: 978-1-59593-593-9. DOI: 10.1145/1240624.1240646.
- Holmes, Andrew (Sept. 2020). “Researcher Positionality - A Consideration of Its Influence and Place in Qualitative Research - A New Researcher Guide”. In: *Shanlax International Journal of Education* 8, pp. 1–10. DOI: 10.34293/education.v8i4.3232.
- Holzinger, Andreas (June 2016). “Interactive machine learning for health informatics: when do we need the human-in-the-loop?” en. In: *Brain Informatics* 3.2, pp. 119–131. ISSN: 2198-4018, 2198-4026. DOI: 10.1007/s40708-016-0042-6.
- Höök, Kristina (Oct. 2010). “Transferring qualities from horseback riding to design”. In: *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*. NordiCHI ’10. New York, NY, USA: Association for Computing Machinery, pp. 226–235. ISBN: 978-1-60558-934-3. DOI: 10.1145/1868914.1868943.
- Hunkenschroer, Anna Lena and Alexander Kriebitz (2023). “Is AI recruiting (un) ethical? A human rights perspective on the use of AI for hiring”. In: *AI and Ethics* 3.1, pp. 199–213.
- Hyesun, Choung, David Prabu, and Ross Arun (2023). “Trust in AI and Its Role in the Acceptance of AI Technologies”. In: *International Journal of Human-Computer Interaction* 39.9, pp. 1727–1739. DOI: 10.1080/10447318.2022.2050543. eprint: <https://doi.org/10.1080/10447318.2022.2050543>.
- Inkpen, Kori et al. (2019). “Where is the Human? Bridging the Gap Between AI and HCI”. In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI EA ’19. Glasgow, Scotland Uk: Association

- for Computing Machinery, pp. 1–9. ISBN: 9781450359719. DOI: 10.1145/3290607.3299002.
- Jacovi, Alon et al. (2021). “Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI”. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. Virtual Event, Canada: Association for Computing Machinery, pp. 624–635. ISBN: 9781450383097. DOI: 10.1145/3442188.3445923.
- Jain, Dhruv et al. (Oct. 2018). “Towards Accessible Conversations in a Mobile Context for People who are Deaf and Hard of Hearing”. In: *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. ASSETS ’18. New York, NY, USA: Association for Computing Machinery, pp. 81–92. ISBN: 978-1-4503-5650-3. DOI: 10.1145/3234695.3236362.
- Jain, Sarthak and Byron C. Wallace (May 2019). *Attention is not Explanation*. arXiv:1902.10186 [cs]. DOI: 10.48550/arXiv.1902.10186.
- Javed, Saleha et al. (Mar. 2021). “Understanding the Role of Objectivity in Machine Learning and Research Evaluation”. en. In: *Philosophies* 6.1, p. 22. DOI: 10.3390/philosophies6010022.
- Jiang, Jinglu et al. (May 2023). “A Situation Awareness Perspective on Human-AI Interaction: Tensions and Opportunities”. In: *International Journal of Human-Computer Interaction* 39.9, pp. 1789–1806. ISSN: 1044-7318. DOI: 10.1080/10447318.2022.2093863.
- Jin, Hanlei et al. (Mar. 2024). *A Comprehensive Survey on Process-Oriented Automatic Text Summarization with Exploration of LLM-Based Methods*. arXiv:2403.02901 [cs]. DOI: 10.48550/arXiv.2403.02901.
- Jootun, Dev, Gerry McGhee, and Glenn R Marland (2009). “Reflexivity: promoting rigour in qualitative research.” In: *Nursing standard* 23.23, pp. 42–47.
- Kapoor, Ashish et al. (2010). “Interactive optimization for steering machine classification”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1343–1352.

- Kim, Been, Rajiv Khanna, and Oluwasanmi O Koyejo (2016). “Examples are not enough, learn to criticize! Criticism for Interpretability”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee et al. Vol. 29. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2016/file/5680522b8e2bb01943234bce7bf84534-Paper.pdf.
- Kim, Been, Martin Wattenberg, et al. (July 2018). “Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)”. In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 2668–2677.
- Kim, Sunnie S. Y. et al. (2023). “”Help Me Help the AI”: Understanding How Explainability Can Support Human-AI Interaction”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. Hamburg, Germany: Association for Computing Machinery. ISBN: 9781450394215. DOI: [10.1145/3544548.3581001](https://doi.org/10.1145/3544548.3581001).
- King, Stephen and Judhi Prasetyo (2023). “Assessing generative AI through the lens of the 2023 Gartner Hype Cycle for Emerging Technologies: a collaborative autoethnography”. In: *Frontiers in Education*. Vol. 8. Frontiers, p. 1300391.
- Kittley-Davies, Jacob et al. (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*. CHI ’19. Glasgow, Scotland UK: Association for Computing Machinery, pp. 1–12. ISBN: 9781450359702. DOI: [10.1145/3290605.3300273](https://doi.org/10.1145/3290605.3300273).
- Koroteev, Mikhail V (2021). “BERT: a review of applications in natural language processing and understanding”. In: *arXiv preprint arXiv:2103.11943*.
- Kuhn, Thomas S (1997). *The structure of scientific revolutions*. Vol. 962. University of Chicago press Chicago.
- Lage, Isaac et al. (Oct. 2019). “Human Evaluation of Models Built for Interpretability”. en. In: *Proceedings of the AAAI Conference on Human Computation and*

- Crowdsourcing* 7, pp. 59–67. ISSN: 2769-1349. DOI: 10.1609/hcomp.v7i1.5280.
- Lai, Vivian and Chenhao Tan (Jan. 2019). “On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* ’19. New York, NY, USA: Association for Computing Machinery, pp. 29–38. ISBN: 978-1-4503-6125-5. DOI: 10.1145/3287560.3287590.
- Lakkaraju, Himabindu and Osbert Bastani (Feb. 2020). “”How do I fool you?”: Manipulating User Trust via Misleading Black Box Explanations”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’20. New York, NY, USA: Association for Computing Machinery, pp. 79–85. ISBN: 978-1-4503-7110-0. DOI: 10.1145/3375627.3375833.
- Langer, Markus et al. (July 2021). “What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research”. In: *Artificial Intelligence* 296, p. 103473. ISSN: 0004-3702. DOI: 10.1016/j.artint.2021.103473.
- Lazer, David et al. (Feb. 2009). “Computational Social Science”. In: *Science* 323.5915, pp. 721–723. DOI: 10.1126/science.1167742.
- Lee, John D. and Katrina A. See (2004). “Trust in Automation: Designing for Appropriate Reliance”. In: *Human Factors* 46.1. PMID: 15151155, pp. 50–80. DOI: 10.1518/hfes.46.1.50_30392. eprint: https://doi.org/10.1518/hfes.46.1.50_30392.
- Lewis, Seth C., Rodrigo Zamith, and Alfred Hermida (Jan. 2013). “Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods”. In: *Journal of Broadcasting and Electronic Media* 57.1, pp. 34–52. ISSN: 0883-8151. DOI: 10.1080/08838151.2012.761702.
- Liao, Ching Ya, Pangfeng Liu, and Jan-Jan Wu (Dec. 2020). “Convolution Filter Pruning for Transfer Learning on Small Dataset”. In: *2020 International Com-*

- puter Symposium (ICS), pp. 79–84. DOI: 10.1109/ICS51289.2020.900025.
- Liao, Q. Vera, Daniel Gruen, and Sarah Miller (2020). “Questioning the AI: Informing Design Practices for Explainable AI User Experiences”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. Honolulu, HI, USA: Association for Computing Machinery, pp. 1–15. ISBN: 9781450367080. DOI: 10.1145/3313831.3376590.
- Liew, Jasy Suet Yan et al. (2014). “Optimizing features in active machine learning for complex qualitative content analysis”. In: *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pp. 44–48.
- Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis (Jan. 2021). “Explainable AI: A Review of Machine Learning Interpretability Methods”. In: *Entropy* 23.1. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute, p. 18. ISSN: 1099-4300. DOI: 10.3390/e23010018.
- Lipton, Zachary C. (Sept. 2018). “The mythos of model interpretability”. In: *Commun. ACM* 61.10, pp. 36–43. ISSN: 0001-0782. DOI: 10.1145/3233231.
- Lucero, Andrés (June 2018). “Living Without a Mobile Phone: An Autoethnography”. In: *Proceedings of the 2018 Designing Interactive Systems Conference*. DIS ’18. New York, NY, USA: Association for Computing Machinery, pp. 765–776. ISBN: 978-1-4503-5198-0. DOI: 10.1145/3196709.3196731.
- Lundberg, Scott M. and Su-In Lee (2017). “A unified approach to interpreting model predictions”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. Long Beach, California, USA: Curran Associates Inc., pp. 4768–4777. ISBN: 9781510860964.
- Luo, Zheheng, Qianqian Xie, and Sophia Ananiadou (Apr. 2023). *ChatGPT as a Factual Inconsistency Evaluator for Text Summarization*. arXiv:2303.15621 [cs]. DOI: 10.48550/arXiv.2303.15621.

- Lyotard, Jean-François (1984). “The postmodern condition: A report on knowledge”. In: *U of Minnesota P.*
- Mackenzie, Adrian (Nov. 2017). *Machine Learners: Archaeology of a Data Practice*. en. MIT Press. ISBN: 978-0-262-03682-5.
- Mainsbridge, Mary (2022). “Feeling movement in live electronic music: An embodied autoethnography”. In: *Proceedings of the 8th International Conference on Movement and Computing*. MOCO '22. Chicago, IL, USA: Association for Computing Machinery. ISBN: 9781450387163. DOI: 10.1145/3537972.3537989.
- Marathe, Megh and Kentaro Toyama (Apr. 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. New York, NY, USA: Association for Computing Machinery, pp. 1–12. ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3173922.
- Mastorakis, Georgios (Nov. 2018). “Human-like machine learning: limitations and suggestions”. In: *arXiv:1811.06052 [cs]*. URL: <http://arxiv.org/abs/1811.06052> (visited on 01/10/2022).
- Medeiros, André Dantas de et al. (July 2020). “Interactive machine learning for soybean seed and seedling quality classification”. en. In: *Scientific Reports* 10.1, p. 11267. ISSN: 2045-2322. DOI: 10.1038/s41598-020-68273-y.
- Miceli, Milagros, Martin Schuessler, and Tianling Yang (Oct. 2020). “Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision”. In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW2, 115:1–115:25. DOI: 10.1145/3415186.
- Mikolov, Tomas et al. (Oct. 2013). *Distributed Representations of Words and Phrases and their Compositionality*. arXiv:1310.4546 [cs, stat]. DOI: 10.48550/arXiv.1310.4546.
- Milana, Federico, Enrico Costanza, and Joel E Fischer (2023). “Chatbots as Advisers: the Effects of Response Variability and Reply Suggestion Buttons”. In: *Proceedings of the 5th International Conference on Conversational User*

- Interfaces*. CUI '23. Eindhoven, Netherlands: Association for Computing Machinery. ISBN: 9798400700149. DOI: 10.1145/3571884.3597132.
- Mohseni, Sina, Jeremy E. Block, and Eric D. Ragan (June 2020). A *Human-Grounded Evaluation Benchmark for Local Explanations of Machine Learning*. arXiv:1801.05075 [cs]. DOI: 10.48550/arXiv.1801.05075.
- Mueller, Shane T. et al. (Feb. 2019). *Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI*. arXiv:1902.01876 [cs]. DOI: 10.48550/arXiv.1902.01876.
- Muller, Michael et al. (Nov. 2016). “Machine Learning and Grounded Theory Method: Convergence, Divergence, and Combination”. In: *Proceedings of the 19th International Conference on Supporting Group Work*. GROUP '16. New York, NY, USA: Association for Computing Machinery, pp. 3–8. ISBN: 978-1-4503-4276-6. DOI: 10.1145/2957276.2957280.
- Muramatsu, Jack and Wanda Pratt (Sept. 2001). “Transparent Queries: investigation users’ mental models of search engines”. In: *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '01. New York, NY, USA: Association for Computing Machinery, pp. 217–224. ISBN: 978-1-58113-331-8. DOI: 10.1145/383952.383991.
- Musgrave, George (2019). “Making sense of my creativity: Reflecting on digital autoethnography”. In: *Journal of Artistic and Creative Education* 13.1, pp. 1–11.
- Nauta, Meike et al. (July 2023). “From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI”. In: *ACM Comput. Surv.* 55.13s. ISSN: 0360-0300. DOI: 10.1145/3583558.
- Neustaedter, Carman and Phoebe Sengers (June 2012). “Autobiographical design in HCI research: designing and learning through use-it-yourself”. In: *Proceedings of the Designing Interactive Systems Conference*. DIS '12. New York, NY,

- USA: Association for Computing Machinery, pp. 514–523. ISBN: 978-1-4503-1210-3. DOI: 10.1145/2317956.2318034.
- Nourani, Mahsan et al. (Oct. 2019). “The Effects of Meaningful and Meaningless Explanations on Trust and Perceived System Accuracy in Intelligent Systems”. en. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, pp. 97–105. ISSN: 2769-1349. DOI: 10.1609/hcomp.v7i1.5284.
- O’Kane, Aisling Ann, Yvonne Rogers, and Ann E. Blandford (Apr. 2014). “Gaining empathy for non-routine mobile device use through autoethnography”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’14. New York, NY, USA: Association for Computing Machinery, pp. 987–990. ISBN: 978-1-4503-2473-1. DOI: 10.1145/2556288.2557179.
- Papenmeier, Andrea, Gwenn Englebienne, and Christin Seifert (July 2019). *How model accuracy and explanation fidelity influence user trust*. arXiv:1907.12652 [cs]. DOI: 10.48550/arXiv.1907.12652.
- Phung, Van Hiep and Eun Joo Rhee (Jan. 2019). “A High-Accuracy Model Average Ensemble of Convolutional Neural Networks for Classification of Cloud Image Patches on Small Datasets”. en. In: *Applied Sciences* 9.21, p. 4500. ISSN: 2076-3417. DOI: 10.3390/app9214500.
- Pijnappel, Sebastiaan and Florian ‘Floyd’ Mueller (Feb. 2014). “Designing interactive technology for skateboarding”. In: *Proceedings of the 8th International Conference on Tangible, Embedded and Embodied Interaction*. TEI ’14. New York, NY, USA: Association for Computing Machinery, pp. 141–148. ISBN: 978-1-4503-2635-3. DOI: 10.1145/2540930.2540950.
- Porter, Reid, James Theiler, and Don Hush (Sept. 2013). “Interactive Machine Learning in Data Exploitation”. In: *Computing in Science Engineering* 15.5, pp. 12–20. ISSN: 1558-366X. DOI: 10.1109/MCSE.2013.74.
- Poursabzi-Sangdeh, Forough, Daniel G Goldstein, et al. (2021a). “Manipulating and Measuring Model Interpretability”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Seattle, WA, USA, May 8–13, 2021, pp. 1–12. ACM, New York, NY, USA, ISBN: 978-1-4503-8108-5. DOI: 10.1145/3411764.3445902.

- ence on Human Factors in Computing Systems.* CHI '21. Yokohama, Japan: Association for Computing Machinery. ISBN: 9781450380966. DOI: 10 . 1145/3411764.3445315.
- Poursabzi-Sangdeh, Forough, Daniel G. Goldstein, et al. (Aug. 2021b). *Manipulating and Measuring Model Interpretability*. arXiv:1802.07810 [cs]. DOI: 10 . 48550/arXiv.1802.07810.
- Rashid, Al M. et al. (Apr. 2006). "Motivating participation by displaying the value of contribution". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* CHI '06. New York, NY, USA: Association for Computing Machinery, pp. 955–958. ISBN: 978-1-59593-372-0. DOI: 10 . 1145 / 1124772.1124915.
- Raychev, Veselin et al. (Jan. 2016). "Learning programs from noisy data". In: *SIGPLAN Not.* 51.1. Place: New York, NY, USA Publisher: Association for Computing Machinery, pp. 761–774. ISSN: 0362-1340. DOI: 10 . 1145 / 2914770.2837671.
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (Aug. 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. New York, NY, USA: Association for Computing Machinery, pp. 1135–1144. ISBN: 978-1-4503-4232-2. DOI: 10 . 1145/2939672.2939778.
- (Apr. 2018). "Anchors: High-Precision Model-Agnostic Explanations". en. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1. Number: 1. ISSN: 2374-3468. DOI: 10 . 1609/aaai.v32i1.11491.
- Riedl, Mark O (2019). "Human-centered artificial intelligence and machine learning". In: *Human behavior and emerging technologies* 1.1, pp. 33–36.
- Riordan, Niamh O (2014). "Autoethnography: Proposing a new method for information systems research". In: *Twenty Second European Conference On Information*. Vol. 22, pp. 1–14.

- Rorty, Richard 1931-2007 (1994). *Consequences of Pragmatism: (essays: 1972 - 1980)*. 6. print. Minneapolis, Minn.: University of Minnesota Press. ISBN: 978-0-8166-1063-1.
- Rosenthal, Stephanie L. and Anind K. Dey (Feb. 2010). "Towards maximizing the accuracy of human-labeled sensor data". In: *Proceedings of the 15th International Conference on Intelligent User Interfaces*. IUI '10. New York, NY, USA: Association for Computing Machinery, pp. 259–268. ISBN: 978-1-60558-515-4. DOI: 10.1145/1719970.1720006.
- Rozenblit, Leonid and Frank Keil (2002). "The misunderstood limits of folk science: an illusion of explanatory depth". en. In: *Cognitive Science* 26.5, pp. 521–562. ISSN: 1551-6709. DOI: 10.1207/s15516709cog2605_1. (Visited on 08/18/2024).
- Rudin, Cynthia (May 2019). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". en. In: *Nature Machine Intelligence* 1.5. Publisher: Nature Publishing Group, pp. 206–215. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0048-x.
- Salih, Ahmed M. et al. (June 2024). "A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME". en. In: *Advanced Intelligent Systems* n/a.n/a, p. 2400304. ISSN: 2640-4567. DOI: 10.1002/aisy.202400304.
- Samek, Wojciech, Thomas Wiegand, and Klaus-Robert Müller (2017). *Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models*. arXiv: 1708 . 08296 [cs.AI]. URL: <https://arxiv.org/abs/1708.08296>.
- Sanchez, Téo et al. (Apr. 2021). "How do People Train a Machine? Strategies and (Mis)Understandings". In: 5.CSCW1. DOI: 10.1145/3449236.
- Sanyal, Soumya and Xiang Ren (Nov. 2021). "Discretized Integrated Gradients for Explaining Language Models". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens et al. Online and Punta Cana, Dominican Republic: Association for

- Computational Linguistics, pp. 10285–10299. DOI: 10.18653/v1/2021.emnlp-main.805.
- Sarkar, Advait, Alan F Blackwell, et al. (July 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)*, pp. 53–56. DOI: 10.1109/VLHCC.2014.6883022.
- Sarkar, Advait, Mateja Jamnik, et al. (Oct. 2015). “Interactive visual machine learning in spreadsheets”. In: *2015 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC’15)*, pp. 159–163. DOI: 10.1109/VLHCC.2015.7357211.
- Scarpato, Noemi et al. (2024). “Evaluating Explainable Machine Learning Models for Clinicians”. In: *Cognitive Computation*, pp. 1–11.
- Schaffer, James et al. (Mar. 2019). “I can do better than your AI: expertise and explanations”. In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI ’19. New York, NY, USA: Association for Computing Machinery, pp. 240–251. ISBN: 978-1-4503-6272-6. DOI: 10.1145/3301275.3302308.
- Schlümm, Jeffrey C. and Richard H. Granger (Sept. 1986). “Incremental learning from noisy data”. en. In: *Machine Learning* 1.3, pp. 317–354. ISSN: 1573-0565. DOI: 10.1007/BF00116895.
- Schneider, Johannes, Christian Meske, and Michalis Vlachos (Dec. 2021). *Deceptive AI Explanations: Creation and Detection*. arXiv:2001.07641 [cs, stat]. DOI: 10.48550/arXiv.2001.07641.
- Severes, Beatriz et al. (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. Orlando, FL, USA: Association for Computing Machinery, pp. 126–132. ISBN: 9798400703362. DOI: 10.1145/3615335.3623062.
- Shen, Hong et al. (Oct. 2020). “Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm

- Performance”. In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW2, 153:1–153:22. DOI: 10.1145/3415224.
- Shneiderman, Ben (2020). “Human-centered artificial intelligence: Three fresh ideas”. In: *AIS Transactions on Human-Computer Interaction* 12.3, pp. 109–124.
- Simard, Patrice Y et al. (2017). “Machine teaching: A new paradigm for building machine learning systems”. In: *arXiv preprint arXiv:1707.06742*.
- Sokol, Kacper and Peter Flach (Jan. 2020). “Explainability fact sheets: a framework for systematic assessment of explainable approaches”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. New York, NY, USA: Association for Computing Machinery, pp. 56–67. ISBN: 978-1-4503-6936-7. DOI: 10.1145/3351095.3372870.
- Srinivasu, Parvathaneni Naga et al. (2022). “From blackbox to explainable AI in healthcare: existing tools and case studies”. In: *Mobile Information Systems* 2022.1, p. 8167821.
- Sternberg, Robert J. and Li-fang Zhang (Apr. 2014). *Perspectives on Thinking, Learning, and Cognitive Styles*. en. Google-Books-ID: YM**e**QAgAAQBAJ. Routledge. ISBN: 978-1-135-66361-2.
- Stolper, Charles D., Adam Perer, and David Gotz (Dec. 2014). “Progressive Visual Analytics: User-Driven Visual Exploration of In-Progress Analytics”. In: *IEEE Transactions on Visualization and Computer Graphics* 20.12, pp. 1653–1662. ISSN: 1941-0506. DOI: 10.1109/TVCG.2014.2346574.
- Stumpf, Simone, Vidya Rajaram, Lida Li, Margaret Burnett, et al. (Jan. 2007). “Toward harnessing user feedback for machine learning”. In: *Proceedings of the 12th international conference on Intelligent user interfaces*. IUI ’07. New York, NY, USA: Association for Computing Machinery, pp. 82–91. ISBN: 978-1-59593-481-9. DOI: 10.1145/1216295.1216316.
- Stumpf, Simone, Vidya Rajaram, Lida Li, Weng-Keen Wong, et al. (2009). “Interacting meaningfully with machine learning systems: Three experiments”. In: *International Journal of Human-Computer Studies* 67.8, pp. 639–662. ISSN:

- 1071-5819. DOI: <https://doi.org/10.1016/j.ijhcs.2009.03.004>.
- Sundar, S Shyam (Jan. 2020). “Rise of Machine Agency: A Framework for Studying the Psychology of Human–AI Interaction (HAII)”. In: *Journal of Computer-Mediated Communication* 25.1, pp. 74–88. ISSN: 1083-6101. DOI: 10.1093/jcmc/zmz026. eprint: <https://academic.oup.com/jcmc/article-pdf/25/1/74/32961171/zmz026.pdf>.
- Teso, Stefano et al. (Feb. 2023). “Leveraging explanations in interactive machine learning: An overview”. English. In: *Frontiers in Artificial Intelligence* 6. Publisher: Frontiers. ISSN: 2624-8212. DOI: 10.3389/frai.2023.1066049.
- Tierney, P.J. (Nov. 2012). “A qualitative analysis framework using natural language processing and graph theory”. In: *The International Review of Research in Open and Distributed Learning* 13, pp. 173–189. DOI: 10.19173/irrodl.v13i5.1240.
- Tsoumakas, Grigorios and Ioannis Katakis (July 2007). *Multi-Label Classification: An Overview*. *International Journal of Data Warehousing and Mining EBSCOhost*. en. ISSN: 1548-3924 Issue: 3 Pages: 1 Volume: 3. DOI: 10.4018/jdwm.2007070101.
- Ueno, Takane et al. (2022). “Trust in Human-AI Interaction: Scoping Out Models, Measures, and Methods”. In: *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI EA ’22. New Orleans, LA, USA: Association for Computing Machinery. ISBN: 9781450391566. DOI: 10.1145/3491101.3519772.
- Väkevä, Jaakko, Elisa D. Mekler, and Janne Lindqvist (2024). “From Disorientation to Harmony: Autoethnographic Insights into Transformative Videogame Experiences”. In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI ’24. Honolulu, HI, USA: Association for Computing Machinery. ISBN: 9798400703300. DOI: 10.1145/3613904.3642543.
- Verame, Jhim Kiel M., Enrico Costanza, and Sarvapali D. Ramchurn (May 2016). “The Effect of Displaying System Confidence Information on the Usage of

- Autonomous Systems for Non-specialist Applications: A Lab Study”. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI ’16. New York, NY, USA: Association for Computing Machinery, pp. 4908–4920. ISBN: 978-1-4503-3362-7. DOI: 10.1145/2858036.2858369.
- Vereschak, Oleksandra, Fatemeh Alizadeh, et al. (2024). “Trust in AI-assisted Decision Making: Perspectives from Those Behind the System and Those for Whom the Decision is Made”. In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI ’24. Honolulu, HI, USA: Association for Computing Machinery. ISBN: 9798400703300. DOI: 10.1145/3613904.3642018.
- Vereschak, Oleksandra, Gilles Bailly, and Baptiste Caramiaux (Oct. 2021). “How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies”. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2. DOI: 10.1145/3476068.
- Waa, Jasper van der et al. (Feb. 2021). “Evaluating XAI: A comparison of rule-based and example-based explanations”. In: *Artificial Intelligence* 291, p. 103404. ISSN: 0004-3702. DOI: 10.1016/j.artint.2020.103404.
- Wall, Emily et al. (Jan. 2018). “Podium: Ranking Data Using Mixed-Initiative Visual Analytics”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1, pp. 288–297. ISSN: 1941-0506. DOI: 10.1109/TVCG.2017.2745078.
- Wallace, Byron C. et al. (Jan. 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. New York, NY, USA: Association for Computing Machinery, pp. 819–824. ISBN: 978-1-4503-0781-9. DOI: 10.1145/2110363.2110464.
- Wang, Danding et al. (2019). “Designing Theory-Driven User-Centric Explainable AI”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI ’19. Glasgow, Scotland Uk: Association for Computing

- Machinery, pp. 1–15. ISBN: 9781450359702. DOI: 10.1145/3290605.3300831.
- Ware, Malcolm et al. (Sept. 2001). “Interactive machine learning: letting users build classifiers”. In: *International Journal of Human-Computer Studies* 55.3, pp. 281–292. ISSN: 1071-5819. DOI: 10.1006/ijhc.2001.0499.
- Waseem, Zeerak et al. (Jan. 2021). *Disembodied Machine Learning: On the Illusion of Objectivity in NLP*. DOI: 10.48550/arXiv.2101.11974.
- Wattenberg, Martin and Fernanda B. Viegas (2008). “The Word Tree, an Interactive Visual Concordance”. In: *IEEE Transactions on Visualization and Computer Graphics* 14.6, pp. 1221–1228. ISSN: 1941-0506. DOI: 10.1109/TVCG.2008.172.
- Wei, Jason et al. (Oct. 2022). *Emergent Abilities of Large Language Models*. arXiv:2206.07682 [cs]. DOI: 10.48550/arXiv.2206.07682.
- Wei Xu Marvin J. Dainoff, Liezhong Ge and Zaifeng Gao (2023). “Transitioning to Human Interaction with AI Systems: New Challenges and Opportunities for HCI Professionals to Enable Human-Centered AI”. In: *International Journal of Human-Computer Interaction* 39.3, pp. 494–518. DOI: 10.1080/10447318.2022.2041900.
- Willig, Carla and Wendy Stainton Rogers (Mar. 2017). *The SAGE Handbook of Qualitative Research in Psychology*. en. SAGE. ISBN: 978-1-5264-2286-6.
- Wobbrock, Jacob O. et al. (May 2011). “The aligned rank transform for nonparametric factorial analyses using only anova procedures”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’11. New York, NY, USA: Association for Computing Machinery, pp. 143–146. ISBN: 978-1-4503-0228-9. DOI: 10.1145/1978942.1978963.
- Wong, Weng-Keen et al. (Feb. 2011). “End-user feature labeling: a locally-weighted regression approach”. In: *Proceedings of the 16th International Conference on Intelligent User Interfaces*. IUI ’11. New York, NY, USA: Association for Computing Machinery, pp. 115–124. ISBN: 978-1-4503-0419-1. DOI: 10.1145/1943403.1943423.

- Xu, Wei (June 2019). “Toward human-centered AI: a perspective from human-computer interaction”. In: *Interactions* 26.4, pp. 42–46. ISSN: 1072-5520. DOI: 10.1145/3328485.
- Yang, Chengliang, Anand Rangarajan, and Sanjay Ranka (May 2018). *Global Model Interpretation via Recursive Partitioning*. arXiv:1802.04253 [cs, stat]. DOI: 10.48550/arXiv.1802.04253.
- Yang, Qian, Aaron Steinfeld, et al. (2020). “Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. Honolulu, HI, USA: Association for Computing Machinery, pp. 1–13. ISBN: 9781450367080. DOI: 10.1145/3313831.3376301.
- Yang, Qian, Jina Suh, et al. (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. Hong Kong, China: Association for Computing Machinery, pp. 573–584. ISBN: 9781450351980. DOI: 10.1145/3196709.3196729.
- Yang, Xianjun et al. (Feb. 2023). *Exploring the Limits of ChatGPT for Query or Aspect-based Text Summarization*. arXiv:2302.08081 [cs]. DOI: 10.48550/arXiv.2302.08081.
- Yu, Liang et al. (Oct. 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)*, pp. 49–56. DOI: 10.1109/VAST.2015.7347630.
- Zhang, Ying and Chen Ling (May 2018). “A strategy to apply machine learning to small datasets in materials science”. en. In: *npj Computational Materials* 4.1, pp. 1–8. ISSN: 2057-3960. DOI: 10.1038/s41524-018-0081-z.