Quality and User Experience (2021) 6:6
https://doi.org/10.1007/s41233-021-00046-5

RESEARCH ARTICLE

Investigating the user experience of customer service chatbot interaction: a framework for qualitative analysis of chatbot dialogues

Asbjørn Følstad[1] · **Cameron Taylor**[2]

Received: 9 December 2020 / Accepted: 13 August 2021 / Published online: 20 August 2021

Abstract
The uptake of chatbots for customer service depends on the user experience. For such chatbots, user experience in particular concerns whether the user is provided relevant answers to their queries and the chatbot interaction brings them closer to resolving their problem. Dialogue data from interactions between users and chatbots represents a potentially valuable source of insight into user experience. However, there is a need for knowledge of how to make use of these data. Motivated by this, we present a framework for qualitative analysis of chatbot dialogues in the customer service domain. The framework has been developed across several studies involving two chatbots for customer service, in collaboration with the chatbot hosts. We present the framework and illustrate its application with insights from three case examples. Through the case findings, we show how the framework may provide insight into key drivers of user experience, including response relevance and dialogue helpfulness (Case 1), insight to drive chatbot improvement in practice (Case 2), and insight of theoretical and practical relevance for understanding chatbot user types and interaction patterns (Case 3). On the basis of the findings, we discuss the strengths and limitations of the framework, its theoretical and practical implications, and directions for future work.

Keywords Chatbot · User experience · Dialogue analysis · Human–chatbot interaction

Introduction

Chatbots are increasingly seen as a valuable complement to customer service [1]. According to a recent Gartner report [2], 31% of interviewed organizations already had, or were in the short-term planning for introducing conversational platforms. Survey data suggests that 40% of retail consumers in the US to have experiences with chatbots [3]. In retail banking and insurance, a leading business sector for the uptake of chatbots, nearly half of the top 100 organizations have implemented chatbots to assist customers [4].

To realize the potential benefits of chatbots for customer service, these should entail positive user experiences [5]. This is particularly important as user uptake of chatbots for customer service has been found to lag behind industry expectations [6]. In consequence, service providers need to strengthen capabilities for assessing and monitoring chatbot user experience –to improve identification and evaluation of key drivers of user experience, provide the needed basis for continuous improvement work, and to gain new insight into chatbot users and their evolving patterns of use. While there exists a substantial body of knowledge on the evaluation of spoken dialogue systems [7], there is a need to strengthen capabilities for assessment and monitoring of user experience in current chatbots for customer service [8] which are typically implemented as text-based solutions within company websites or messaging applications [6]. In the present study, we consider qualitative analysis of chatbot dialogues as a means to provide needed capabilities for assessment and monitoring of factors relevant to user experience in such chatbots.

Chatbot dialogues are potentially valuable sources of insight into user experience. In such dialogues, users express their service requests in their own words—typically as an initial question or enquiry to the chatbot which may be refined through subsequent turn-taking between the user and the chatbot. As such, the dialogues may provide insight into user experience, issues in service provision, and emerging user needs [9]. However, the research leveraging such chatbot dialogues as a source of insight typically extends only to the

✉ Asbjørn Følstad
asf@sintef.no

[1] SINTEF, Oslo, Norway

[2] boost.ai, Sandnes, Norway

use of automated analysis approaches [10–12], which entails the risk of overlooking aspects of the conversation that would be observable to a human analyst. Only a few studies present qualitative analysis of chatbot dialogue data [13–15], and there is limited guidance on how to conduct such analysis. This dearth of studies is problematic as it suggests an unrealized potential in the utilization of dialogue data as a resource for improving user experience in practical chatbot development and maintenance. Also, it suggests an opportunity for theory development needed for improving user experience in chatbots. Furthermore, a common framework for the analysis of chatbot dialogues would enable comparison for purposes of practice and research.

In response to this, we present a framework for qualitative analysis of dialogues in customer service chatbots. The framework mainly concerns pragmatic aspects of user experience as these have been shown to be of particular importance to current chatbots for customer service [16, 17]. Following an iterative development process, in collaboration with chatbot case partners, the framework provides an overarching structure for analysing dialogue data at the level of turn-taking between the user and chatbot and at the level of entire dialogues, with *response relevance*, *understandability*, *dialogue helpfulness*, and *dialogue efficiency* as key constructs.

To demonstrate its application, as well as its benefits and limitations, we present three case examples where the framework has been applied in the analysis of chatbot data along with related implications and lessons learnt. As such, this paper provides two main contributions. First, the presented framework provides needed support for research on chatbot dialogue data and for chatbot development and improvement practice in service providers. Second, the presented case applications of the framework provide insight into key drivers of user experience as well as how chatbot dialogue data may support practical improvement work as well as theory building.

The structure of the remainder of the paper is as follows. First, we provide an overview of relevant background and present the research objective, key requirements for the framework, and the research context. We then present the framework development process and the framework itself, before presenting three case examples where the framework has been applied. Finally, we discuss the framework relative to the requirements and background literature, consider its limitations, and suggest future work.

Background

The user experience of chatbots for customer service

Customer service is a key application domain for chatbots [4]. Such chatbots are typically text-based and task-oriented [9]. A range of platforms are available for chatbots. These typically provide natural language processing capabilities to identify user intents from free-text input as well as facilities for analytics [2]. Users will likely interact with such chatbots by presenting their requests in free text, as they would when interacting with human customer service personnel. When continuing the dialogue, users may do this through subsequent free-text responses or by using buttons or quick replies to navigate the intent hierarchy [18]. As such, current chatbots for customer service typically combine what McTear [7] refers to as statistical data-driven systems based on machine learning and rule-based systems that are handcrafted using best practice guidelines. Analytic capabilities provided for such chatbots may include metrics of traffic from users, which intents users trigger and how often, failures in the chatbot to make intent predictions, as well as features for user ratings or feedback. Improvement efforts in chatbots for customer service typically include updating the intent hierarchy and associated content, adjusting the training data driving intent predictions, and adding integrations with backend systems [9].

We understand user experience as concerning users' perceptions and responses from use and anticipated use of interactive systems [19], driven by factors such as pragmatic and hedonic quality, beauty, and goodness [20]. Users' motivation for engaging with chatbots in general [21] and chatbots for customer service in particular [5], is primarily to get efficient and effective access to information or support. Hence, the pragmatic quality of customer service chatbots is highly important [14, 16, 18]. This is not to say that non-pragmatic aspects are irrelevant. For example, introducing anthropomorphic cues in chatbots for customer service have been found to have positive implications for user compliance with a chatbot's feedback requests [6] and emotional connection with the company following chatbot interaction [22]. However, given the prominence of pragmatic aspects of user experience for customer service chatbots, addressing such quality in user experience assessments is seen as critical for chatbots to be taken up more broadly by users for customer service purposes [1, 5]. For example, van der Goot et al. [17], in a recent study of customer journeys that involve chatbots for customer service, found that companies' top priority in improving chatbot user experience should be efficient and effective problem resolution. Hence, in our framework, we see drivers of customer experience associated with pragmatic quality as particularly important.

Analysis of chatbot dialogue

The quality of chatbot conversations is key to good chatbot user experience [23]. Chatbot users note the importance of conversational intelligence and the importance of chatbots being able to retain conversational context [24].

Furthermore, studies have shown the potential impact of message interactivity in chatbots [25], the importance of conversational flow [26, 27] and conversational repair [28]. Also, studies suggest the importance of adapting conversation to fit the user type [15, 29]. In spite of this, it has been argued that quality in chatbot conversational design has not kept pace with the developments in the uptake of such chatbots [14]. To strengthen conversational design, conversation analysis is used to guide the design [29] and redesign [30] of chatbot conversations.

Given the relative importance of conversational design, and how the intended chatbot conversations actually play out in the meetings between chatbots and users, it is critical to leverage data from chatbot conversations for insight into user experience. Recent studies have used chatbot dialogue data for automated analysis approaches, such as text mining [10], sentiment analysis [11], and log analysis [12]. Qualitative approaches to analysis of such data are also presented to some extent. Li et al. [14] conducted an inductive analysis of conversations between users and a banking chatbot for insight into non-progress and coping strategies in the chatbot dialogues. Liao et al. [15] used qualitative analysis to investigate user interaction patterns for a company internal chatbot, identifying differences between social and utilitarian users.

Frameworks developed and used for analysis of interaction with other types of chatbots and conversational agents are potentially relevant also to customer service chatbots. Drawing on a literature review of quality assessment in chatbots and conversational agents, Radziwill and Benton [8] proposed a generic framework for assessment of quality attributes such as performance, functionality, human likeness, affective appeal, and accessibility. In the domain of social chatbots, an area of research with increased current research interest following the availability of large language models [31, 32], Adiwardana et al. [13] proposed a metric for assessing the sensibleness and specificity of chatbot responses and applied this on Google Meena. Previously, Hill et al. [33] and Lortie and Guitton [34] have applied the Linguistic Inquiry and Word Count (LIWC) framework to analyse dialogues between social chatbots and users.

Of higher relevance for customer service chatbots is previous research from the domain of spoken dialogue systems, given the relatively strong pragmatic and task-oriented character typically bestowed on such systems. Within this research tradition, there is a preference for automated evaluation of spoken dialogue systems [7], for example by way of next utterance classification [35] or by way of error simulations [36], where evaluations may involve analysis on message sequence level and dialogue level. However, the much-cited PARADISE framework [37] support analysis of user satisfaction also by means of qualitative interpretation of dialogue data. Here, dialogues with an agent are assessed in terms of task completion, efficiency, and other aspects of relevance to user satisfaction such as the ratios of inappropriate utterances and repair. This framework has also been sought extended by use of dialogue act tagging [38] for additional insight into dialogue quality and improved prediction of user satisfaction. However, the specific characteristics of chatbots for customer service, including text-based interaction, a large number and variety of intents and associated actions, links between the chatbot and associated online content, as well as the option of escalation to human personnel, arguably require an approach more specifically fit to this type of application than what is provided in the PARADISE framework.

Finally, conversation analysis [39] provides a generic framework for descriptive analysis of dialogue which is of relevance to chatbot evaluation. Here, dialogue is seen as a series of turns between speakers with implicit rules for speakership and turn-taking [40]. Interactions between conversational partners follow a pattern of adjacency pairs, such as inquiry-answer or offer-accept/reject, which may be expanded to form longer sequences. The universals for human–human conversation described in conversation analysis are considered applicable also in the design for, and analysis of, human–machine conversation [41]. These universals support assessment by providing structure to chatbot dialogues on which assessment constructs may be applied. In our framework, presented below, the concept of message sequence [40] identifies the unit of assessment for message relevance and understandability.

Method

Research approach

To leverage chatbot dialogues as a source of insight into user experience, our research objective was to establish a framework for qualitative analysis of such data. Specifically, we aimed for the framework to address the most important current aspects of user experience in such chatbots, that is, aspects concerning effectiveness and efficiency. The framework was particularly intended to support initial qualitative analyses of chatbot dialogues and thereby serve as a basis for further qualitative and quantitative explorations or assessments.

To establish the framework, our development process followed a design science research approach [42]. This is a suitable approach for research aiming to create and evaluate artifacts that help solve problems in the real world—i.e., our framework for qualitative analysis of dialogues with chatbots for customer service. In this method section, we detail the requirements for the framework, the research context, and the development process.

Requirements for the framework

The framework was developed in response to a real world need to enable systematic analysis of dialogue data from customer service chatbot in order to gain insight into customer experience. To guide the development and assessment of the framework, a set of requirements were detailed. The requirements are presented in Table 1.

Research context

The context of the framework development was a collaborative research and innovation project involving, among others, two case companies using a chatbot for customer service, a chatbot platform provider, and a research organization.

The two case companies were part of the project by virtue of their experience and expertise in applying chatbots for customer services. The companies had both applied chatbots for customer service for several years. Their chatbots are implemented as part of their online presence, available through the companies' customer websites interact with several thousand users monthly.

The chatbot platform provider was part of the project as a leading provider with high level expertise in chatbots for customer service. It hosts chatbots for more than 100 client companies and, as such, has substantial experience concerning the analysis needs and requirements of their clients.

The research organization was part of the project by virtue of their expertise on user-centred design and user needs and requirements for chatbot interactions.

Development process

The framework was developed in collaboration with the partners of the collaborative project and validated and refined across several studies. The purpose of these studies was to strengthen insight into user experience and to provide knowledge needed for further development and refinement of the chatbot solutions.

The development of the framework followed an iterative process involving two versions of the framework. The development process ran in the period 2018–2020, including (a) initial explorations of analysis approaches, (b) interim framework development and application, (c) final framework development and application.

Initial explorations of analysis approaches

Initial explorations of analysis approaches were conducted on anonymous chatbot dialogue datasets from two case companies (in total, 1910 dialogues). The aim of the initial explorations was to consider a range of aspects in the dialogue data with respect to their relevance for user experience assessment. Specifically, we analysed messages and dialogues with respect to types of user requests, user motivations for contacting the chatbot, dialogue processes and dialogue outcomes, as well as dialogue topics. During the initial explorations, we were sensitized to the challenge of false positives and false negatives in chatbot responses. Furthermore, we were sensitized to the phenomenon of a small but significant number of dialogues reflecting users' failing to understand the chatbot as a machine and users' experiencing usability issues. When presenting findings to case companies, we found results concerning dialogue process and -outcome to be seen as particularly useful. However, our analyses concerning types of user requests, user motivations for contacting the chatbot and topics of the conversations were found to be less helpful for understanding and improving chatbot user experience and, therefore, less relevant to pursue.

Interim framework development and application

The interim framework development and application were conducted on the basis of experiences from the initial explorations. The interim framework was developed in a workshop involving one of the case companies, the chatbot platform provider and the research partner, and later verified with the second case company. In the interim framework,

Table 1 Requirements for the framework for analysis of dialogue data from chatbots for customer service

Requirements	Details
R1: Insight into drivers of user experience	The framework should, on the basis of analysis of dialogue data, enable insight into key drivers of user experience in chatbots for customer services
R2: Establish benchmarks	The framework should enable establishment of quantifiable benchmarks for user experience on the basis of qualitative analysis of dialogue data
R3: Enable monitoring and comparison of performance	The framework should enable quantifiable monitoring and comparison of chatbot user experience over time and across instances on the basis of qualitative analysis of dialogue data
R4: Allow for general application across chatbots	The framework should be sufficiently general so as to be applicable for different customer service chatbots

we included several aspects for analysis at the levels of turn-taking between the user and chatbot and entire dialogues. Aspects included prediction accuracy, dialogue conclusiveness, dialogue helpfulness, usability issues, out-of-scope requests, and dialogue directedness. The interim framework was applied in a preliminary analysis of chatbot dialogues with one of the case companies [43] including 700 chatbot dialogues. Based on the feedback from the case company, the interim framework was found relevant but somehow complex due to the number and partial overlap of analysis aspects.

Final framework development and application

The development of the current version of the framework, and its application in the presented case examples, drew on the experiences from application of the interim framework. In response to the challenge of complexity and overlap, the aspects of the framework were condensed and restructured by the research partner and presented to the other project partners for feedback. The framework is presented in fourth section and its application is presented in the three case examples of fifth section.

Framework

In this section we first provide an overview of chatbot dialogues as a data source, before detailing the framework constructs concerning dialogue turn-taking and entire dialogues respectively. In Sect. 5, we then show applications of the framework.

Chatbot dialogues as data for qualitative analysis—key terms

The data source when analysing chatbot dialogues are the messages exchanged between the chatbot and the user, the metadata for these messages, and information on user interaction with interactive elements in the chat dialogue.

By dialogue, we mean an interaction session between a user and the chatbot. Within dialogues, messages from the user and the chatbot are structured in smaller clusters. We refer to such clusters as message sequences. In line with Schegloff [40], the basic message sequence is an adjacency pair where the user and chatbot contribute one message each that are related to each other. However, message sequences may also include expansions to the adjacency pair, in the form of pre-expansions, inserted expansions, and post-expansions.

A dialogue consists of one or more message sequences. A new message sequence is typically started when a message from the user is to be interpreted as a new intent, not just an expansion of the ongoing sequence. Within message sequences, users often are provided interactive elements such as buttons, quick replies, or links to external content. Figure 1 shows examples of message sequences within dialogues. Here, we also provide an overview of key framework constructs introduced later in this section.

The importance of privacy protection needs to be accentuated when engaging in qualitative analysis of chatbot dialogues. This is particularly relevant as chatbot dialogues include messages from users in free text. In our work, we have resolved this by analysing anonymous dialogues. Furthermore, all free text has been automatically scanned for content in the free text that may inadvertently identify users. Any such content has been masked in accordance with established data processing agreements between the case partners and the research organization.

Analysis of message sequences—response relevance and understandability

User experience during a chatbot dialogue develops through message sequences, where the user and chatbot take turns. Hence, analysis of each message sequence is important. Drawing on the experiences of our framework development process, we see two constructs as particularly useful in such analysis: Response relevance and understandability.

Response relevance

The user experience of a message sequence depends on whether the chatbot response is a relevant reply to the user's message. A key challenge in current chatbots for customer service is that intent predictions may return false positives, leading the chatbot to give irrelevant replies. However, the chatbot may also fail to identify any intent in the user's message, potentially leading to a false negative where the chatbot will not provide a relevant answer although this is available in its knowledge base. It is therefore essential to assess the quality of chatbot responses to user messages.

We apply four categories for initial qualitative analysis of chatbot responses: *relevant response*, *false positive*, *false negative*, and *out of scope*. The categories are detailed and exemplified in Table 2 below. Note that all examples are constructed for the purpose of explanation, not drawn from the later presented cases.

Analysis of response relevance echo the qualitative measures of the PARADISE framework [37]. Specifically, the category *false positive* resembles the PARADISE concept *inappropriate utterance* and the category *out of scope* partially overlap the PARADISE concept *repair*; the latter because *out of scope* responses typically are provided as repairs. However, the categories of our framework are particularly adapted to chatbots for customer service. Hence,

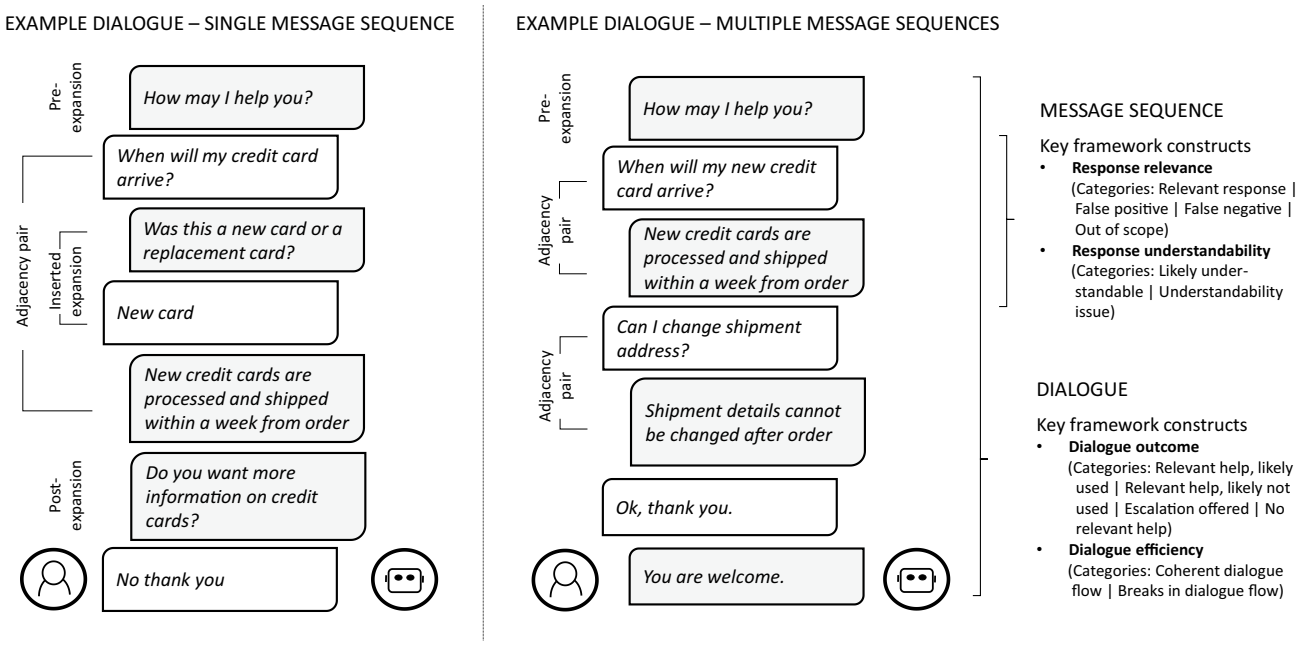

Fig. 1 Example chatbot dialogues illustrating different message sequences. To the right are key framework constructs listed

the categories are chosen so as to address the chatbot's success in predicting users' intents rather than the response of the dialogue system. This, because the main challenge in customer service chatbots is not to provide a relevant answer once a user intent has been identified, but to predict users' intents from their free text input.

Response understandability

While the response may be relevant, the user experience also depends on the user understanding the response from the chatbot. Challenges in this regard may include failure to understand the content of the response or failure to understand available interactive elements such as buttons and quick replies. For qualitative analysis of response understandability, we apply a binary categorization: *likely understandable* and *understandability issue*. The categories are detailed and exemplified in Table 3.

Analysis of entire dialogues—outcome and efficiency

While analysis at the level of message sequences provide insight into the relevance and value of individual chatbot responses, analysis at the level of entire dialogues provides insight into the interaction outcome as well as the user effort spent in achieving this. On the basis of our experiences from the framework development process, we consider two

Table 2 Categories for analysing response relevance, with descriptions and dialogue examples

Category	Description	Example
Response relevance		
Relevant response	The response is relevant for the user message	User: When will my new credit card arrive? Chatbot: I can help you order a new credit card
False positive	The response is irrelevant for the user message	User: When will my new credit card arrive? Chatbot: I can help you order a new credit card
False negative	The response erroneously indicates the user message to be out of scope	User: When will my new credit card arrive? Chatbot: Sorry, I do not have an answer. You may try to rephrase your question
Out of scope	The response correctly indicates the user message to be out of scope	User: What is your favourite ice cream flavour? Chatbot: Sorry, I do not have an answer. You may try to rephrase your question

constructs to be particularly useful: Dialogue outcome and dialogue efficiency.

Dialogue outcome

The outcome of the dialogue concerns whether the user has received needed support. A qualitative analysis of dialogue outcome requires assumptions regarding the user's intended goal for the chatbot interaction. Unless other data sources are available, such as users self-reported data on problem resolution, these assumptions will be based on an interpretation of the user messages during the dialogue—in particular, the initial formulation of the service request. While it may seem a daunting task to interpret user goals based on their own presentations of these as messages to a chatbot, our experience is that such assumptions arguably may be made with reasonable confidence.

As the first step in the analysis of dialogue outcomes, we apply the following categories: *relevant help likely used*, *relevant help likely not used, escalation offered*, *no relevant help*. The categories are detailed and exemplified in Table 4.

Dialogue efficiency

Efficiency in dialogue concerns the perceived time and effort required by the user to achieve the desired outcome. Analysing dialogue efficiency is in our experience not straight forward as the number of message sequences required to reach a conclusion of the dialogue may not be a good indication of perceived effort. Rather, a dialogue with several message sequences that gradually produces a more precise understanding of the user's problem may be seen as efficient as a dialogue with fewer such sequences. However, in line with findings from the literature, dialogues where users have to rephrase requests, or have repeated message sequences with the chatbot with little seeming progress, are likely perceived as inefficient. We address this in the qualitative analysis by identifying breaks in the dialogue flow, that is, message sequences that divert the dialogue from its intended goal or where there is no progress towards the users intended goal—for example when the chatbot fails to interpret the users request. A main benefit of this analysis lies in discriminating between dialogue with such breaks in the flow and dialogues without, as such breaks often indicate user experience issues—for example due to interpretation failure. In consequence, we apply the following binary categorization: *coherent dialogue flow*, *breaks in dialogue flow*. The categories are detailed and exemplified in Table 5.

Analysis of dialogue outcome and dialogue efficiency resemble key concepts of the PARADISE framework [37]. Specifically, *dialogue outcome* resembles the PARADISE concept *task success* and *dialogue efficiency* is related to the PARADISE efficiency measures. However, whereas in PARADISE task outcome and efficiency are quantitative measures, the concepts in our framework presuppose qualitative assessment. In customer service chatbots, this is necessary as the user goal may not be immediately evident from the system data about the interaction—due to the problem of false positives. Furthermore, from a user experience perspective, dialogue efficiency may not be quantified only as the number of interactions, but rather as the breaks in the dialogue flow, as interactions leading consistently towards a goal arguably are perceived differently by users than interactions not leading towards the intended goal. For example, dialogues going in circles, not leading towards the intended goal, is considered a key user experience issue in customer service chatbots [5].

Summary of framework and possible next steps in analysis

The presented framework consists of constructs at two levels of analysis: the level of message sequences and the level of entire dialogues. For each level of analysis, two

Table 3 Categories for analysing response understandability, with descriptions and dialogue examples

Category	Description	Example
Response understandability		
Likely understand-able	The message sequence and its dialogue context does not indicate understandability issues	User: When will my new credit card arrive? Chatbot: New credit cards are processed and shipped within a week from order. To check order status, please click here. [Button with link] User: [Clicks button with link]
Understand-ability issue	The message sequence and its dialogue context, such as users' post-expansions or failure to use interactive elements, indicate understandability issues	User: When will my new credit card arrive? Chatbot: New credit cards are processed and shipped within a week from order. To check order status, please click here. [Button with link] User: How can I check status?

Table 4 Categories for analysing dialogue outcome, with descriptions and dialogue examples

Category	Description	Example
Dialogue outcome		
Relevant help likely used	The chatbot offers support corresponding to the assumed user goal, either directly in the chat or through interactive elements used by the user	User: How can I check the status of my order of a new credit card? Chatbot: To check the order status of your card, please click here. [Button with link] User: [Clicks button with link]
Relevant help likely not used	The chatbot offers support corresponding to the assumed user goal through interactive elements, but this is not taken advantage of by the user	User: How can I check the status of my order of a new credit card? Chatbot: To check the order status of your card, please click here. [Button with link] [End of conversation]
Escalation offered	The chatbot offers escalation to other means of service, for example manned customer service, due to interpretation failure or to company policy	User: There is an error in my last invoice Chatbot: Sorry, I do not have an answer. You may try to rephrase your question. You can also ask for help from one of my human colleagues
No relevant help	The chatbot offers insufficient help corresponding to the assumed user goal. No escalation offered	User: There is an error in my last invoice Chatbot: Sorry, I do not have an answer. You may try to rephrase your question [End of conversation]

Table 5 Categories for analysing dialogue efficiency, with descriptions and dialogue examples

Category	Description	Example
Dialogue efficiency		
Coherent dialogue flow	The dialog contains no breaks due to misinterpretations or failure to advance towards the user's goal	User: When will my new credit card arrive? Chatbot: New credit cards are processed and shipped within a week from order. To check order status, please click here. [Button with link] User: That's ok. Thanks Chatbot: You're welcome
Breaks in dialogue flow	The dialogue contains breaks due to misinterpretations or failure to advance towards the user's goal	User: When will my new credit card arrive? Chatbot: I can help you order a new card User: No, I have already ordered. When will I receive it? Chatbot: Sorry, I do not have an answer. You may try to rephrase your question

key constructs are detailed, each with corresponding coding categories as shown in Tables 2, 3, 4 and 5.

The framework is intended to serve the purpose of understanding drivers of chatbot user experience. Though key drivers are already identified in the framework, it may be relevant for researchers or practitioners to explore also other possible drivers of user experience and also to investigate other aspects of user types, preferences, and behaviour. As such, the framework may also be applied as a first step in a more in-depth analysis.

Following initial analysis of message sequences, sequences coded as including false positives, false negatives or understandability issues may be made subject to inductive qualitative analysis to identify problematic patterns and potential resolutions. This is exemplified in Case 1, presented in the section "Case 1: analysis to gain insight into key drivers of user experience".

Following initial analysis of entire dialogues, dialogues coded as relevant help not used, escalation offered, or no relevant help—as well as dialogues with breaks—may be made subject to inductive qualitative analysis to identify possible causes and potential resolutions. This is exemplified in Case 3, presented in the section "Case 3: analysis to support exploratory insights of relevance to user experience".

Case experiences

To exemplify the application of the proposed framework, we present experiences from three case applications. The three cases are chosen so as to show the benefit and limitations of the framework, and how the framework output may enable further qualitative and quantitative explorations.

Case 1: analysis to gain insight into key drivers of user experience

Case 1 demonstrates how the framework may be applied to gain insight into key drivers of user experience, in particular the pragmatic quality of the interaction. We provide detail on how all constructs of the framework were applied, the resulting findings, and the implications of these for further analysis and potential for improvement.

Background

The framework was applied for an analysis of users' interactions with a customer service chatbot for a financial service company. The customer service chatbot could identify and respond to several thousand user intents related to the service domain and was offered as a source of support through the company customer website.

The analysis was conducted without other data available than the chatbot dialogues. The chatbot dialogue data were anonymous and included the sequence of messages between the user and the chatbot, metadata on time and predicted intent for each message as well as information on the use of buttons or links provided by the chatbot.

Method

678 chatbot dialogues were randomly sampled from the set of user dialogues with the chatbot for a given week of November 2019. The sampled dialogues were evenly distributed across the days of the week so as to avoid bias due to variations across the week. A small subset of dialogues was filtered out of the sample, including dialogues for testing the chatbot, dialogues with no meaningful request, dialogues of one particular service category (requests for a particular form), and immediate requests for human personnel.

The dialogues were analysed manually according to the presented framework. For each dialogue, the first message sequence containing a user request was coded for response relevance (relevant response, false positive, false negative, out of scope) and understandability (understandability issues—yes/no). Furthermore, each dialogue was coded for dialogue helpfulness (relevant help likely used, relevant help not used, escalation offered, no relevant help) and efficiency (coherent dialogue flow—yes/no). Following the initial dialogue analysis, findings of particular relevance for insight into user experience were summarized.

Findings

The analysis of dialogue message sequences indicated that the chatbot was able to predict and respond to the intents reflected in user messages for most of the user requests. A total of 66% of the responses were coded as relevant, whereas 5% were coded as out of scope. False positives were more prevalent (23%) than false negatives (6%) suggesting that for at least some intents the prediction threshold may be biased towards providing a response rather than the default fallback. Furthermore, 9% of the dialogues were found to suggest understandability issues in the text of provided interaction mechanisms, indicating that the responses provided by the chatbot mostly were easy to process by the users. Also, false positives were typically not found detrimental to understandability. Users receiving false positive responses often chose to continue their dialogue, typically rephrasing their initial request as an attempt of recovery.

The analysis of dialogues found that in 36% of the dialogues, help was offered and likely used whereas in another 46%, escalation to manned customer service was offered. About half the escalations were due to company policy in certain product and service areas. In 16% of the dialogues no help was offered and in 2% help was offered but not used. Dialogues were overall found to be efficient, with only 21% of the dialogues identified as having breaks in the dialogue flow. However, it should be noted that most dialogues (57%) only had one message sequence.

Implications and lessons learnt

The findings hold several noteworthy implications. First, the findings exemplify how the framework provide insight into key drivers of user experience for customer service chatbots, in particular chatbot response relevance and dialogue helpfulness. In particular, the findings suggest the importance of false positives to user experience, as nearly a quarter of all message sequences include a false positive. The findings also strongly suggest the benefit to user experience of providing escalation from the chatbot to manned personnel. Nearly half the conversations included offers of escalation to manned customer service; escalations that likely are important for the resulting user experience. The findings may also serve as basis for further qualitative explorations, for example to identify types of intents likely to return false positive responses or to understand characteristics of dialogues not perceived as helpful.

Lessons learnt from the analysis process include that analysis of chatbot responsiveness and dialogue helpfulness seem feasible, though the evaluator needs to make assumptions regarding users' true intents based only on the text provided. A challenging aspect of the analysis is the interpretation of dialogues abandoned by the users prior to these receiving or using relevant help. It is difficult to know the users' reasons for abandoning dialogues, as this could be due to either dissatisfaction with the chatbot responses or interruption of the interaction for reasons outside the chatbot in the user context.

Case 2: analysis to support benchmarking and comparison

Case 2 concerns how the framework may support practical efforts in improving a chatbot for customer service. The system change concerned the introduction of uncertainty responses, were the chatbot in cases of low prediction confidence expressed uncertainty regarding its interpretation of the user request and suggested different possible means of support. The case analysis was intended to (a) identify how the change had impacted user experience, and (b) provide guidance for further improvements in the chatbot.

Background

In Case 2, the framework was applied in a before–after study. The analysis was conducted in a similar manner to that of Case 1, though in two waves to allow for comparison between the two chatbot versions. The case domain also for Case 2 was customer service for financial services. Preliminary findings from the comparative study have been presented previously [43]. Here, we present findings from the complete analysis following the most recent version of the analysis framework and also include findings of relevance to subsequent chatbot improvement work.

The chatbot in Case 2 was built on the same underlying platform as the chatbot in Case 1, though in a different case company, and was offered as a source of support through the company customer website. In the analysis, we assessed the chatbot according to the framework constructs for benchmarking and comparison purposes.

The analysis was conducted without other data available than the chatbot dialogues. The chatbot dialogue data were anonymous and included the sequence of messages from the user and the chatbot for each dialogue, metadata on time and predicted intent for each message as well as information on use of buttons or links provided by the chatbot.

Method

1400 chatbot dialogues were randomly sampled from the set of user dialogues with the chatbot for two subsequent weeks in 2019, 700 each week before and after the change in the chatbot. The dialogues were sampled to be evenly distributed across the days of both weeks. The following types of dialogues were not included in the sample: dialogues for testing the chatbot, dialogues with no meaningful request, and dialogues concerning a small number of atypical service categories.

The dialogues were made subject to initial qualitative analysis according to the presented framework in the same way as in Case 1. Following the initial dialogue analysis, findings were summarized to allow for comparing of the two chatbot versions. Findings were further analysed for specific groups of user intents in the chatbot to guide future development. All analyses were conducted manually.

Findings

As in Case 1, the analysis of message sequences indicated that the chatbot was able to predict and respond to the intents reflected in user messages for most of the user requests. Before implementation of the change, 57% of the responses were coded as relevant and 12% coded as out of scope, and similar numbers were found after the change. The comparative analysis, however, showed a significant benefit of the new version in terms of a marked reduction in false positives from 28% before the change to 14% after, a statistically significant reduction ($\chi^2 = 64.2$, $p < 0.001$). This reduction in false positive responses corresponded to the emergence of a new set of message sequences providing an uncertainty response.

The analysis of complete dialogues demonstrated findings that were comparable to Case 1: In 30% of the dialogues help was offered and likely used and in 33% escalation to manned customer service was offered. For the latter group of dialogues, about half the escalations were due to company policy in certain product and service areas—similar to what was found in Case 1, though the two case companies do not have identical policies for escalation.

Interestingly, no marked differences were observed between the two chatbot versions in Case 2 with regard to the entire dialogue process: The proportion of dialogues where help was offered and likely used was practically unchanged (30% pre implementation, 29% post implementation; $\chi^2 = 0.1$, $p = 0.73$), through there was a tendency towards reduction in dialogues with no help offered (25% pre implementation and 21% post implementation; $\chi^2 = 3.6$, $p = 0.06$) and a tendency towards increase in dialogues where escalation was offered (22% pre implementation and 27% post implementation; $\chi^2 = 4.2$, $p = 0.07$). Possibly, the tendency to a reduction in dialogues in with no help offered may be directly related the tendency to an increase in dialogues where escalation was offered, though the analysis does allow for a firm conclusion on this.

The analysis findings were also broken down for the most frequent user intents. In total, eighteen intents of the more than 400 intents predicted in the sample dialogues, had enough sample dialogues for detailed analysis. These eighteen intents included nine high-level intents and nine intents further down in the intent hierarchy. This additional analysis helped distinguish between intents in terms of prevalence of false positives prior to and after the introduction of the new chatbot version. Among the most prevalent intents, five were found to have above 30% false positives in chatbot replies also after implementing the new version of the chatbot. This

insight was useful to guide subsequent improvement work as it helped prioritize resources among groups of intents. Since the five intents in question were relevant for more than 10% of the sampled dialogues, improvements in these may substantially impact user experience.

Implications and lessons learnt

The findings demonstrate the relevance of the framework for practical improvement work in chatbots. First, the framework supports assessments of the effect of a major change in the chatbot on user experience, the introduction of functionality for the chatbot to expressing uncertainty and offer alternatives to the user. It was demonstrated that the major benefit of the change was to reduce the number of false positives in the chatbot, and thereby counter one of the key issues users typically report regarding chatbots for customer service [5]. At the same time, it was noted that while the effect on individual message sequences was substantial, the effect on the level of entire dialogues was smaller—something that suggests that users find ways to adapt to false positives in chatbot responses. Second, the framework was found to provide insight that may guide improvement work by identifying intents and intent groups in which there may be prediction issues, something that is highly useful information to guide updating of training data in the chatbot.

Lessons learnt from the analysis process include the relative ease with which qualitative analysis based on the framework lent itself to purposes of benchmarking and comparison. It may also be noted that such comparison would not have been feasible on the basis of analytics based on automatically predicted intents in the chatbot. Such automated analysis would not enable the identification of false positives which, in the analysis based on the framework, was found to be important for improvement work on the chatbot.

Case 3: analysis to support exploratory insights of relevance to user experience

Case 3 is included as the third case example to show the potential value of the framework to support exploratory user research of theoretical and practical significance. Here, we applied the analyses conducted in Case 2 as basis for investigating differences among chatbot user groups, specifically between user groups applying different modes of communication with the chatbot: more business-like or more social. The analysis from Case 2 was used for this purpose by conducting additional analysis of user messages in the chatbot dialogues—in terms of these including signifiers of socially oriented, specifically the use of greetings and first- and second person pronouns.

Background

During analyses we noted that while some users are highly formal and business-like in their communication with the chatbot for customer service, others are more socially oriented, that is, personal and informal. These two modes of communication with a chatbot have interesting characteristics, where the former may be briefer and more to-the-point and the latter may be more detailed and subjectively oriented. The research literature seems surprisingly silent on distinguishing between these two modes in users chatbot communication, though previous work has analysed differences in how users communicate with chatbots and humans [33] and how users adapt their conversation when using voice-based conversational agents [44]. An exception to this is a study by Liao et al. [15] on conversational search in a company internal chatbot. Here, they found evidence for utilitarian and social chatbot user types corresponding closely to the findings in our study.

Method

Taking as basis the analysis presented in Case 2, we added an analysis of the users' messages in the chatbot dialogue. Also this analysis was conducted manually. Echoing the suggested markers of social orientation by Liao et al. [15], we analysed whether users' messages included (a) greetings and displays of politeness (e.g., "hi" and "thank you") and (b) uses of first- and second person pronoun (e.g., "I need help with …", "Can you help me with …"). Dialogues containing both (a) and (b) were categorized as socially-oriented. Socially- and non-socially oriented dialogues were compared with regard to response relevance for the message sequences and helpfulness for the entire dialogues. Following this, inductive analyses of socially-oriented dialogues were conducted to understand why and how there were differences between the groups. The analyses presented in our study is preliminary. The complete analyses will be the subject of a future publication.

Findings

In our dataset, 28% of the dialogues were found to reflect users with a social orientation whereas 72% reflected users with a non-social orientation. The analysis suggested significant differences in dialogue process and outcome between the user groups. Users with a non-social orientation were found to more often receive relevant responses from the chatbot (67% vs. 52%, significant at $p < 0.01$ in a Fischer's exact test) and more often to receive help within in the chatbot dialogue (39% vs. 24%, significant at $p < 0.01$ in a Fischer's exact test). To understand why and how there was such a difference between the user groups, we conducted an

inductive analysis of selected dialogues. While preliminary, these analyses suggest that possible reasons for increased chance of receiving false positives and lowered chance of receiving help in the chatbot dialogue, may be in part that users with a social orientation tend to overestimate the conversational intelligence of the chatbot—assuming a dialogue that may gradually unfold across many message sequences—and in part that users with a social orientation may provide more rich and detailed messages making intent prediction more challenging.

Implications and lessons learnt

The analyses of Case 3 exemplify the potential theoretical benefits to be drawn from applying the framework. In this case, the findings suggest new insight into different user groups, where their mode of communicating with the chatbot may affect their user experience. Specifically, the findings may complement existing knowledge, in part by pointing out the relevance of exploring dialogue data for evidence of different patterns or styles of interacting with chatbots (cf. [33]). The findings extend the work of Liao et al. [15] by showing the distinction between utilitarian and socially oriented chatbot interaction to be relevant also for the domain of customer service. Furthermore, the findings contribute insight into the implications of this distinctions and also allow us to take steps towards explaining these implications.

Lessons learnt in Case 3 include an understanding of the benefit of additional analyses to the constructs provided in the framework. Hence, it will be important to update, extend and improve on the framework going forward.

Summarizing experiences across cases

Summarizing our case experiences from validating the proposed framework, an overview of the cases and key findings are presented in Table 6. This overview serves as basis for our discussion on benefits and limitations of the framework in terms of the identified framework requirements.

Discussion

In this section, we discuss our presented framework relative to the requirements and consider benefits and limitations based on the presented case examples. Finally, we propose directions for future work.

A framework for qualitative analysis of chatbot dialogues—benefits and limitations

The presented framework is intended as a conceptual structure for qualitative analysis of chatbot dialogues, to support

Table 6 Summary of case results and experiences

Case	Validation	Details	Findings and implications	Relevant requirements
1	Whether framework generates insight into key drivers of user experience	Case analysis of key framework concepts, at message sequence level and dialogue level	Findings show distribution of message sequences and dialogues across coding categories. Provide insight into key drivers of user experience	Validation particularly relevant for R1
2	Whether framework supports benchmarking and comparison	Case analysis using framework concepts to benchmark user experience and apply this for comparison	Findings show changes in key drivers of user experience across two chatbot instances. Provide insight into usefulness for benchmarking and comparison	Validation particularly relevant for R2 and R3
3	Whether framework supports exploratory user experience insight	Case analysis using framework to explore relation between user behaviour and user experience	Findings show how framework concepts may be used to distinguish user behaviour of relevance for user experience, namely social vs. non-social orientation	Validation particularly relevant for R1

practice and theory formation. Identified needs within industry and research led us to outline four requirements for such analysis support. In the following, we discuss our framework relative to these requirements.

Insight into key drivers of user experience (R1)

Previous work shows that pragmatic aspects of user experience are key in chatbots for customer service [5, 16]. Such chatbots are typically highly task- and goal oriented [14, 18] and efficient and effective request resolution is highly important to users [17]. The case experiences from applying our framework suggest that the framework provides valuable insight for pragmatic aspects of user experience, both at the message sequence level—addressing chatbot response relevance and understandability—and at the level of entire dialogues—addressing dialogue outcome and -efficiency.

At the level of message sequences, a particular benefit of the framework is its accentuation of false positives, as this has been found to be among the main causes of unsuccessful chatbot dialogue [9] and a reason for users not to deepen their engagement with chatbot [45]. The problem of false positives is particularly challenging for customer service chatbots as these typically are intended to identify and support a broad range of user intents, something that may increase the risk of the chatbot to make erroneous predictions. Both chatbots in the presented cases included several thousand intents, something that implies a substantial challenge of intent prediction. The rates of false positives in the cases further accentuate the importance of qualitative analysis of response relevance in general and false positives in particular.

Likewise, for analysis of entire dialogues, application of the framework constructs concerning dialogue outcome and process provide valuable insight to practitioners and researchers alike—as shown by the case examples. The promise of efficiency, effectiveness and convenience has been identified as main motivations for chatbot use in general [21] and the pragmatic quality of the dialogue outcome and process is also the premier determinants of user experience in chatbots for customer service [16, 17]. Hence, the constructs of dialogue helpfulness and dialogue efficiency are valuable to the assessment of user experience in this type of chatbots. This is not to say that automated analyses may not generate highly relevant insight for chatbot dialogue outcomes, for example in terms of conversion metrics. However, for customer service chatbots, much of the outcome of chatbot dialogue is provided through the textual or verbal content of the chatbot—for example by the chatbot providing information or guidance—without the user providing feedback on whether or not the outcome was helpful. Hence, qualitative analysis may serve as a valuable complement to automated analyses.

It may also be noted that the concepts and categories of our framework are closely related to concepts of previous frameworks. Specifically, key concepts of the PARADISE framework for analysis of spoken dialogue systems [37] resemble concepts in the presented framework both at the level of the entire dialogue and at the level of message sequences. At dialogue level, the PARADISE concept of *task success* resembles the concept of *dialogue outcome*, and the PARADISE concept of *efficiency* resembles the concept of *dialogue efficiency*. A key difference, though, is that the qualitative analysis of the presented framework allows for more nuanced insight, something that is particularly important when the risk of erroneous interpretation of user intent is high—as is the case in chatbots for customer service given the large volumes of available intents and the challenge of free text interpretation. Likewise, at message sequence level, the PARADISE concepts of *inappropriate utterance* and *repair* resemble the categories *false positive* and *out of scope*. The PARADISE concepts address mainly the system response, whereas our framework concerns the message sequence. This is, in particular, seen in the differentiation between the categories *false negative* and *out of scope*. For both these categories of message sequences, the chatbot will typically return a repair message. However, in the case of a false negative the repair is due to failure to prediction an existing intent, whereas in the *out of scope* message sequence the repair is due to the user going outside the intended chatbot scope.

Automated analysis clearly is important to evaluation of conversational systems, including chatbots; particularly for reasons of cost and efficiency [7]. However, automated analysis may not be able to identify issues of high importance to the user experience of chatbots for customer service, such as false positives. In consequence, a framework for qualitative analysis may serve as a needed complement to existing frameworks.

The presented framework was also shown to provide insight into drivers of user experience beyond assessments of relevance to efficiency and goal achievement. In Case 3 we observed how the analysis supported by the constructs of the platform could also provide insight into the potential impact of individual differences in chatbot interaction style—in part by adding other constructs to the analysis (that of socially vs. non-socially oriented interaction) and in part by providing a basis for further inductive explorations of qualitative data for the identified categories. Hence, while pragmatic quality is the starting point of analyses based on the presented framework, explorations may lead to identification of other factors relevant to user experience. And while non-pragmatic aspects, such as human likeness and language style of the chatbot, likely are of lesser importance than pragmatic aspects for customer service chatbots [17] it will be important for service providers to consider also these

in assessments—in particular as future maturing chatbots for customer service will likely compete on such non-pragmatic aspects once an acceptable level of pragmatic quality has been achieved.

Support for benchmarking (R2) and monitoring and comparison of performance over time (R3)

Support for benchmarking and monitoring and comparison of performance over time is of paramount importance in chatbot development and improvement. The importance of benchmarking and comparison is shown in a recent study by Kvale et al. [8], where analysis of chatbot performance was used to prioritize chatbot intents for improvement work—guiding AI training resources based on the frequency of chatbot intents being triggered and the performance of these relative to other, better performing intents.

The usefulness of the framework for benchmarking and comparison was demonstrated in Case 1 and 2. In Case 1, chatbot performance of relevance for user experience in terms of pragmatic quality was quantified based on the qualitative analysis. Such quantification arguably lends itself to benchmarking purposes across or within chatbot implementations. For example, one could envision performance indexes based on combining response relevance and dialogue helpfulness. Such performance measures would be more credible than comparing, for example, fallback rate—a measure while easily gathered in automated analysis is vulnerable to variation in false positives in the chatbot.

In Case 2, the quantified chatbot performance, based on key framework constructs, was compared for two versions of a chatbot to understand how an implemented change had impacted user experience. This comparison demonstrated a reduction in false positives—that is, a marked effect on response relevance—while at the same time showing only limited effect on dialogue outcome. Such capabilities for benchmarking, monitoring and comparison may be valuable, for example, for assessing different approaches to chatbot informational content, conversational design, or interaction mechanisms. By supporting such assessments, the framework potentially will strengthen future chatbot design and development. At the same time, it may be noted that the manual character of the analysis process implies an inherent limitation in the framework, in that assessments will have to be conducted on relatively small samples of dialogue data. While such manual analysis fits well with current approaches to improvement of training data and intent prediction, future work may address automated analysis support for the framework constructs. At the same time, automated identification of, for example, false positives is inherently challenging and suggests a need also for human oversight of the analysis process in the foreseeable future.

The generality of the framework (R4)

The generality of the framework is likely substantial as its key constructs arguably are relevant for chatbot dialogues across current chatbot platforms used for customer service. Also, the constructs of the framework, addressing message sequences and entire dialogues, are sufficiently generic so as to be compatible with constructs for more nuanced analysis from conversation analysis (e.g., [41]). This said, the case experiences were only gathered for two chatbots, both built on the same chatbot platform. Furthermore, the cases did not include studies integrating the constructs from the framework with constructs based on, for example, linguistic theory [14]. Hence, future research is needed to verify framework generality and compatibility with relevant theoretical approaches.

When discussing the generality of the framework, it is also important to note that the framework likely is of less relevance to chatbots that are not highly task-oriented. Other frameworks and constructs may be needed for analysis of chatbot dialogue in less goal-oriented chatbots, such as the open domain chatbot Google Meena [13] or chatbots for mental health [46]. Here, constructs such as sensibleness and specificity [13], constructs addressing domain-specific success criteria, such as level of self-disclosure for the mental health domain [47], or constructs addressing hedonic aspects of chatbot interaction may be of particular relevance. Future research is needed on frameworks to support qualitative analysis of chatbot dialogue for other domains than customer service.

Our discussion of the four requirements has served to point out several strengths of the framework, including its capacity for benchmarking and monitoring, its potential for providing needed insight into user experience and its assumed generality. However, the framework also has important weaknesses. Among these, is the resource requirements associate with running qualitative analyses, which clearly are higher than when relying on frameworks for automated analysis [7] or the analytics facilities included in chatbot platforms. In response to this weakness, we argue that the resource requirements for analyses applying the framework needs to be balanced against the potential practical benefits of the analysis; in particular, the degree to which an analysis would enable significant improvements to chatbot responses or dialogue outcomes. Another weakness may be a possible lack of flexibility in the framework to adapt to future developments in chatbot platforms, as the framework proposes specific constructs (e.g., response relevance) rather than, for example, proposing a process of reaching needed constructs for a given assessment context. In response to this weakness, we argue a need to adapt and extend the framework over time to keep up with the current state of the art in chatbot platforms and implementations. Finally, while the framework

provides substantial support for analysis of drivers for pragmatic user experience, less direct support is provided for analysis of drivers for other aspects of user experience, such as hedonic quality. This limitation suggests a potential need for future research and will be addressed below.

Study limitations and future work

The framework and case experiences may motivate several directions for future work. In part, we foresee such future work to address limitations in our study. In part, we foresee the framework to motivate future research on user and conversational characteristics as well as user needs and experience.

The main limitations of our presented study lie in the characteristics of the current applications and case examples, as well as the intended scope of the framework. The cases were conducted with a small sample of customer service chatbots within a particular time and geography. Hence, we have not verified the applicability of the framework in a broad range of contexts and future research clearly is needed to verify the adequacy of the framework also for other customer service chatbots, across geographies, and also across time. Furthermore, as the framework is intended for use in the context of chatbots for customer service, future research is needed to establish such frameworks also for other chatbots domains, specifically domains that are less task-oriented.

Future work on the framework will also need to reflect that chatbots for customer service is an application area which has seen rapid developments over the last few years, and likely will see substantial and fast-paced developments also in the near future. Such developments likely will concern both the maturity of the user population and the availability of increasingly advanced chatbots. For example, Gartner predicts that a substantial proportion of chatbots for customer service currently implemented will be abandoned over the coming few years [2], likely replaced by more advanced providers and solutions. For such a changing application area, the framework likely will need repeated validation across case contexts.

When applying the framework in new case contexts at different geographies and times, we foresee that researchers will identify useful new constructs or a need to adapt current constructs—for example to incorporate constructs from conversation analysis [39]. In particular, as chatbots for customer service mature and different solutions do not differentiate strongly with respect to pragmatic quality, the framework may need to be strengthened in its coverage of constructs of relevance to other aspects of user experience. For example, a future framework may benefit from including constructs addressing identity cues and conversational cues in the chatbot [25]. Hence, we consider the presented framework to have an initial character and foresee its future developments to fit the changing needs of the rapidly evolving field of chatbot research and design.

In spite of the initial character of our framework, and the limitations in the presented study, our current case experiences induce optimism in terms of the potential of the framework to motivate new research within the emerging field of human-chatbot interaction. In consequence, we complete our discussion on future work by pointing out what we see as directions for research and practice motivated by the opportunities provided by the framework.

For future research we foresee the following directions for application of the framework as particularly promising:

Understanding the dynamics of chatbot conversations

There is a need for knowledge on how variations in user behaviour and conversational design affect user experience in chatbot interaction. We foresee the framework being used to assess effects of such variations, which will be potentially useful to—for example—identify effects of different approaches to onboarding users to the chatbot, suggest opportunities or service offers to users, or manage conversational repair. Strengthening such knowledge, may ultimately improve chatbot conversational intelligence, accentuated as important to user experience [24].

Understanding variations in user type

Here we foresee studies that address how different user types imply different user behaviours, which in turn generates different user experience. In Case 3 we saw how user interaction style variations impacted the pragmatic quality of interaction. Future work may seek to identify additional or alternative user types and behavioural patterns, the impact of this variation on user experience, and also how individuals may evolve their user type characteristics and behaviour across time and contexts. Knowledge of user types will be valuable to strengthen chatbot adaptation to users needs, preferences, and possibly even personality [29].

For future chatbot development practice, we foresee the following directions for utilizing the framework:

Integrating framework in practice

To benefit from the framework in chatbot development and maintenance, it needs to be embedded in the practices of developers and hosts. Specifically, processes for efficiently utilizing the framework in the context of chatbot design and maintenance needs to be established. As part of such

integration, it will be valuable to draft and implement guidelines noting, for example, the frequency and comprehensiveness of routine assessments as well as the communication and use of findings for benchmarking and comparison. Future reports on integrations of the framework will be valuable addition to the body of knowledge for chatbot practitioners.

Framework to complement automated analysis

We also foresee then need to establish practical integrations of the presented framework for qualitative analysis and available facilities for automated analysis. For example, simple processes for using qualitative analysis may be to check and verify assumptions drawn from automated analysis, and to use qualitative analysis to identify aspects of chatbot interaction of particular relevance for automated analysis.

Conclusion

In this paper we have presented a framework for qualitative analysis of chatbot dialogue in the context of customer service. We have detailed the development process leading to the framework and its validation across three case studies. The validation suggests that the framework may provide insight into key drivers of user experience in chatbots for customer service, and also may be useful for purposes of benchmarking and comparison. Future work is needed to assess the generality of the framework across a broader range of chatbots.

We believe the framework will be a valuable basis for future research and practice, in particular as it provides a way to benefit from chatbot dialogues as a resource for gathering new knowledge on chatbot user experience and also as a means for practical improvement in existing chatbots for customer service. As such, we hope the presented framework may contribute to strengthening the uptake and benefit users may have from chatbots for customer service.

Funding This work was conducted as part of the innovation project Chatbots for Loyalty, supported by the Research Council of Norway, grant no. 282244. Open access funding provided by SINTEF AS.

Declarations

Conflict of interest The second author was an employee of the company providing the chatbot platform used by the case companies in the study.

References

1. Nordheim CB, Følstad A, Bjørkli CA (2019) An initial model of trust in chatbots for customer service—findings from a questionnaire study. Interact Comput 31(3):317–335
2. Gartner (2019) Market guide for virtual customer assistants. Technical report, Gartner. https://www.gartner.com/en/documents/3947357/market-guide-for-virtual-customer-assistants
3. Statista (2020) Share of consumers who have used chatbots to engage with companies in the United States as of 2019, by industry. Statistics report. https://www.statista.com/statistics/1042604/united-stated-share-internet-users-who-used-chatbots-industry/
4. Taylor MP, Jacobs K, Subrahmanyam KVJ et al (2020) Smart talk. How organizations and consumers are embracing voice and chat assistants. Technical report. CapGemini
5. Drift (2018) The 2018 State of chatbots report. Technical report, Drift
6. Adam M, Wessel M, Benlian A (2020) AI-based chatbots in customer service and their effects on user compliance. Electron Mark. https://doi.org/10.1007/s12525-020-00414-7
7. McTear M (2021) Conversational AI: dialogue systems, conversational agents, and chatbots. Morgan & Claypool
8. Kvale K, Freddi E, Hodnebrog S, Sell OA, Følstad A (2020) Understanding the user experience of customer service chatbots: what can we learn from customer satisfaction surveys? In: Proceedings of the CONVERSATIONS 2020 international workshop on chatbot research. Springer, Cham, pp 205–218
9. Kvale K, Sell OA, Hodnebrog S, Følstad A (2019) Improving conversations: lessons learnt from manual analysis of chatbot dialogues. In: Proceedings of the CONVERSATIONS 2019 international workshop on chatbot research. Springer, Cham, pp 187–200.
10. Akhtar M, Neidhardt J, Werthner H (2019) The potential of chatbots: analysis of chatbot conversations. In: IEEE conference on business informatics (CBI). IEEE, pp 397–404
11. Feine J, Morana S, Gnewuch U (2019) Measuring service encounter satisfaction with customer service chatbots using sentiment analysis. In: Proceedings of the 14th international conference on Wirtschaftsinformatik—WI2019. AISeL, pp 1115–1129
12. Jalota R, Trivedi P, Maheshwari G, Ngomo ACN, Usbeck R (2019) An approach for ex-post-facto analysis of knowledge graph-driven chatbots—the DBpedia chatbot. In: Proceedings of the CONVERSATIONS 2019. Springer, Cham, pp 19–33
13. Adiwardana D, Luong MT, So DR, Hall J, Fiedel N, Thoppilan R et al (2020) Towards a human-like open-domain chatbot. arXiv preprint arXiv:2001.09977
14. Li CH, Yeh SF, Chang TJ, Tsai MH, Chen K, Chang YJ (2020) A conversation analysis of non-progress and coping strategies with a banking task-oriented chatbot. In: Proceedings of CHI 2020, paper no. 82. ACM, New York

15. Liao QV, Geyer W, Muller M, Khazaen Y (2020) Conversational interfaces for information search. Understanding and improving information search. Springer, Cham, pp 267–287
16. Følstad A, Skjuve M (2019) Chatbots for customer service: user experience and motivation. In: Proceedings of the 1st international conference on conversational user interfaces—CUI 2019, paper no. 1. ACM, New York
17. van der Goot MJ, Hafkamp L, Dankfort Z (2020) Customer service chatbots: a qualitative interview study into customers' communication journey. In: Proceedings of CONVERSATIONS 2020 international workshop on chatbot research. Springer, Cham, pp 190–204
18 Shevat A (2017) Designing bots: creating conversational experiences. O'Reilly Media, Boston
19. ISO (2010) Ergonomics of human–system interaction—part 210: human-centred design for interactive systems. International Standard. ISO, Geneva
20. Law ELC, Van Schaik P (2010) Modelling user experience—an agenda for research and practice. Interact Comput 22(5):313–322
21. Brandtzaeg PB, Følstad A (2017) Why people use chatbots. In: Proceedings of INSCI 2017. Springer, Cham. pp 377–392
22. Araujo T (2018) Living up to the chatbot hype: the influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. Comput Hum Behav 85:183–189
23. Hall E (2018) Conversational design. A Book Apart, New York
24. Jain M, Kumar P, Kota R, Patel SN (2018) Evaluating and informing the design of chatbots. In: Proceedings of the designing interactive systems conference—DIS 2018. ACM, New York, pp 895–906
25. Go E, Sundar SS (2019) Humanizing chatbots: the effects of visual, identity and conversational cues on humanness perceptions. Comput Hum Behav 97:304–316
26. Gnewuch U, Morana S, Adam M, Maedche A (2018) Faster is not always better: understanding the effect of dynamic response delays in human-chatbot interaction. In: Proceedings of the European conference on information systems—ECIS2018. AISeL, paper no. 113
27 Skjuve M, Haugstveit IM, Følstad A, Brandtzaeg PB (2019) Help! Is my chatbot falling into the uncanny valley? An empirical study of user experience in human–chatbot interaction. Hum Technol 15(1):30–54. https://doi.org/10.17011/ht/urn.201902201607
28. Ashktorab Z, Jain M, Liao QV, Weisz JD (2019) Resilient chatbots: repair strategy preferences for conversational breakdowns. In: Proceedings of CHI 2019, paper no. 254. ACM, New York
29. Ruane E, Farrell S, Ventresque A (2020) User perception of text-based chatbot personality. In: Proceedings of the CONVERSATIONS 2020 international workshop on chatbot research. Springer, Cham, pp 32–47
30. Dippold D, Ladiynden J, Shrubsall R, Ingram R (2020) A turn to language: how interactional sociolinguistics informs the redesign of prompt: response chatbot turns. Discourse Context Media. https://doi.org/10.1016/j.dcm.2020.100432
31. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P et al (2020) Language models are few-shot learners. arXiv preprint arXiv:2005.14165
32. Roller S, Dinan E, Goyal N, Ju D, Williamson M, Liu Y et al (2020) Recipes for building an open-domain chatbot. arXiv preprint arXiv:2004.13637
33. Hill J, Ford WR, Farreras IG (2015) Real conversations with artificial intelligence: a comparison between human–human online conversations and human–chatbot conversations. Comput Hum Behav 49:245–250
34. Lortie CL, Guitton MJ (2011) Judgment of the humanness of an interlocutor is in the eye of the beholder. PLoS ONE 6(9):e25085
35. Lowe R, Serban IV, Noseworthy M, Charlin L, Pineau J (2016) On the evaluation of dialogue systems with next utterance classification. arXiv preprint arXiv:1605.05414
36. Möller S, Englert R, Engelbrecht K, Hafner V, Jameson A, Oulasvirta A et al (2006) MeMo: towards automatic usability evaluation of spoken dialogue services by user error simulations. In: Ninth international conference on spoken language processing
37. Walker MA, Litman DJ, Kamm CA, Abella A (1997) PARADISE: a framework for evaluating spoken dialogue agents. arXiv preprint cmp-lg/9704004
38. Walker M, Passonneau R (2001) DATE: a dialogue act tagging scheme for evaluation of spoken dialogue systems. AT&T Labs-Research, Atlanta
39. Sacks H, Schegloff E, Jefferson G (1974) A simplest systematics for the organization of turn-taking for conversation. Language 50(4):696–735. https://doi.org/10.2307/412243
40. Schegloff EA (2000) Overlapping talk and the organization of turn-taking for conversation. Lang Soc 29(1):1–63
41 Moore RJ, Arar R (2019) Conversational UX design: a practitioner's guide to the natural conversation framework. Morgan & Claypool, Williston
42 Hevner AR, March ST, Park J, Ram S (2004) Design science in information systems research. MIS Q 28:75–105
43. Følstad A, Taylor C (2019) Conversational repair in chatbots for customer service: the effect of expressing uncertainty and suggesting alternatives. In: Proceedings of CONVERSATIONS 2019. Springer, Cham, pp 201–214
44. Luger E, Sellen A (2016) "Like having a really bad PA" the gulf between user expectation and experience of conversational agents. In: Proceedings of CHI 2016. ACM, New York, pp 5286–5297
45. Forrester (2016) The state of chatbots. Technical report, Forrester. https://www.forrester.com/report/The+State+Of+Chatbots/-/E-RES136207
46. Fitzpatrick KK, Darcy A, Vierhile M (2017) Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 4(2):e19
47. Ho A, Hancock J, Miner AS (2018) Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. J Commun 68(4):712–733