DESIGNING A CONVERSATION MINING SYSTEM FOR CUSTOMER SERVICE CHATBOTS

Research Paper

Daniel Schloß, Karlsruhe Institute of Technology, Germany, daniel.schloss@kit.edu

Juan David Gutierrez Espitia, Karlsruhe Institute of Technology, Germany, j.gutierrez@hsag.info

Ulrich Gnewuch, Karlsruhe Institute of Technology, Germany, ulrich.gnewuch@kit.edu

Abstract

As chatbots are gaining popularity in customer service, it is critically important for companies to continuously analyze and improve their chatbots' performance. However, current analysis approaches are often limited to the question-answer level or produce highly aggregated metrics (e.g., conversations per day) instead of leveraging the full potential of the large volume of conversation data to provide actionable insights for chatbot developers and chatbot managers. To address this challenge, we developed a novel chatbot analytics approach — conversation mining — based on concepts and methods from process mining. We instantiated our approach in a conversation mining system that can be used to visually analyze customer-chatbot conversations at the process level. The results of four focus group evaluations suggest that conversation mining can help chatbot developers and chatbot managers to extract useful insights for improving customer service chatbots. Our research contributes to research and practice with novel design knowledge for conversation mining systems.

Keywords: *Chatbots, Customer Service, Conversation Mining, Process Mining, Business Intelligence & Analytics, Design Science Research*

1 Introduction

Driven by technological improvements in natural language understanding (NLU; Dale, 2016), many companies are now deploying automated voice- and conversation-based assistance systems, so-called conversational agents (McTear et al., 2016). Due to the general popularity of chat and instant messaging and the fact that conversational agents are ideal for uniform routine processes, text-based chatbots in particular have become popular in customer service (De Keyser et al., 2019; Gnewuch et al., 2022). They support customers with general inquiries or specific processes such as an address search or a cancellation, combining the advantages of parallelization and permanent availability with a dialog-based experience for the user that has a low barrier to entry (Brandtzaeg and Følstad, 2018, McTear et al., 2016; Schuetzler et al., 2021). In particular, the customer service of large companies receiving a large number of uniform B2C requests, such as those in finance, banking, insurance, energy, or public administration, can benefit from self-service automation, leading to a predicted increase in chatbots of +169% until 2026 (Juniper Research, 2022). However, the potential benefits of chatbot technology for businesses can only be realized if the chatbot is adopted by a wide range of customers (Bordoloi et al., 2021; Gao et al., 2021). As research has shown, customer service chatbot users primarily value fast and functional processing (Brandtzaeg and Følstad, 2018). Chatbots must therefore be good at 1) correctly identifying users' intentions and 2) processing their request (Beaver and Mueen, 2020). To ensure this performance of a customer service chatbot in use, continuous monitoring and optimization by chatbot managers and developers are needed, as neglecting chatbots ("bot rot") lowers their quality (Brandtzaeg

and Følstad, 2018; Følstad and Taylor, 2021; Janssen et al., 2021). Inperformant chatbots, in turn, lead to frustration, perception of poor service and avoidance (Diederich et al., 2021; Riquel et al., 2021).

There are several approaches to inform chatbot managers and developers about possibilities for chatbot improvement. For example, user tests or interviews with chatbot or industry experts can be used as a basis for identifying potentials (e.g., Brandtzaeg and Følstad, 2017; Huang and Chueh, 2021). These, however, can be affected by the (experimental) setup, represent a smaller and possibly skewed sample and require manual work (Kvale et al., 2020). For these reasons, the field of chatbot analytics has become increasingly established in practice and in chatbot research (e.g., Beaver and Mueen, 2020; Følstad and Taylor, 2021). Chatbot analytics relies on the (growing) amount of log data of customer chatbot interactions and designs methods to explore these large amounts of implicit feedback aiming to guide chatbot developers and managers to useful insights about their chatbots. So far, however, the field of chatbot analytics has strongly focused on performance and problems on the single user message (utterance) level (Yaeli and Zeltyn, 2021). For example, a user chatbot conversation can break down if the user's intention (the so-called „intent") is not correctly recognized, i.e., classified by the NLU (Følstad and Taylor, 2021; Rozga, 2018). Nevertheless, problems with chatbots often occur or become apparent in the progress of a conversation (Beaver and Mueen, 2020). Additionally, established metrics that do not refer to the utterance level are often on a too high level (e.g. the number of interactions) and thus do not capture conversational processes and progress sufficiently (Przegalinska et al., 2019).

One discipline dedicated to process-level data analysis is process mining, which analyzes patterns in sequences of discrete steps (Van der Aalst et al., 2012). Therefore, we use the methods of process mining and apply it to the implicit processes of logged customer chatbot conversations. In doing so, we address the stated need for (design) knowledge for chatbot analytics to analyze the flow and progress of conversations (Yaeli and Zeltyn, 2021). In this paper, we present the development of this new chatbot analytics method, which we call "conversation mining". Furthermore, we develop a corresponding (conversation mining) system. In doing so, we follow the design science research (DSR) paradigm to ensure a high degree of user-centeredness in addition to scientific rigor (Gregor and Jones, 2007). Furthermore, following the suggestions of Hevner (2007), our DSR project was conducted jointly with an industry partner, which improves the relevance of the research. Thus, our DSR project is dedicated to answering the following research question:

RQ: *How to design a conversation mining system to support chatbot developers and managers in analyzing customer-chatbot conversations to improve chatbot performance?*

According to the research question, our goal is to address the research gap and user needs (i.e., of chatbot developers and managers) for easily accessible, specifically visual, chatbot analytics (Yaeli and Zeltyn, 2021). We aim to equip chatbot developers and managers with instruments that reduce the need for manual analysis and supplement high level metrics with more detailed insights on chatbot performance on a process level (Kvale et al., 2020). In this paper, we report on the first design cycle (DC) of our DSR project, dedicated to the development of our method and the artifact. For this, in the following second chapter on related work, we first introduce basic terms and knowledge about chatbots, chatbot analytics and process mining. In chapter 3, we explain our methodological DSR approach. The individual stages of our project, from problem space to the final evaluation of our proposed solution, are then presented in detail in chapter 4. In chapter 5, we conclude with a discussion and an outlook on future research.

2 Related Work

2.1 Chatbots in Customer Service

In customer service, so-called „task-focused chatbots" are used (Schuetzler et al., 2021; Grudin and Jacques, 2019). The conversations with those chatbots are rather short and functional and performance aspects are important to chatbot users (De Keyser et al., 2019; Van der Goot et al, 2021). The typical process of a customer-chatbot conversation is shown in **Figure 1**. First, customers have an abstract intention (1), which they express in „utterances" (2) (Kucherbaev et al., 2018). Then, for retrieval-based

chatbots currently used in customer service, artificial intelligence (i.e. NLU) classifies the utterance to an intent, whereupon a corresponding chatbot response (4a) or several fixed dialog steps (4b) are retrieved from a database orchestrated by dialog management (Kucherbaev et al., 2018, Rozga, 2018).

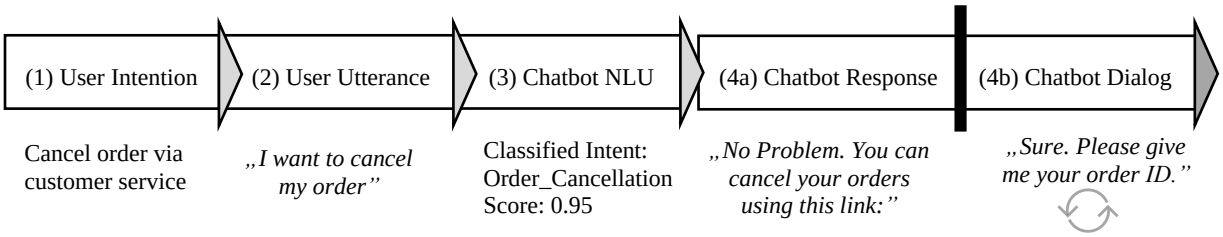


Figure 1. Typical Conversation with a Customer Service Chatbot (Følstad and Taylor, 2021)

Within this process, the first potential source of problems in the conversation is the user expressing his/her concern linguistically imprecise or incorrect (2). Given a user has properly expressed his/her concern, the success of the conversation depends on the NLU capabilities of the chatbot (3) (Schuetzler et al., 2021). The conversation may break down ("Sorry, I did not understand you..") if a topic is out of scope or the NLU is insufficiently trained (Følstad and Taylor, 2021; Galitsky, 2019), ultimately negatively affecting the satisfaction with the conversation as well as the overall service (Diederich et al., 2021; Van der Goot et al, 2021). If a user's intent was correctly identified, the chatbot's response can be insufficient or inappropriate (4a) (Følstad and Taylor, 2021). Finally, problems can occur during guided dialogs with multiple user inputs or question-answers turns (4b), for example, when a chatbot connected to ERP backends requires certain inputs or formats, which the customer does not submit, for completing a conversation (Kucherbaev et al., 2018; Meyer von Wolff et al., 2019; Stolcke et al., 2000).

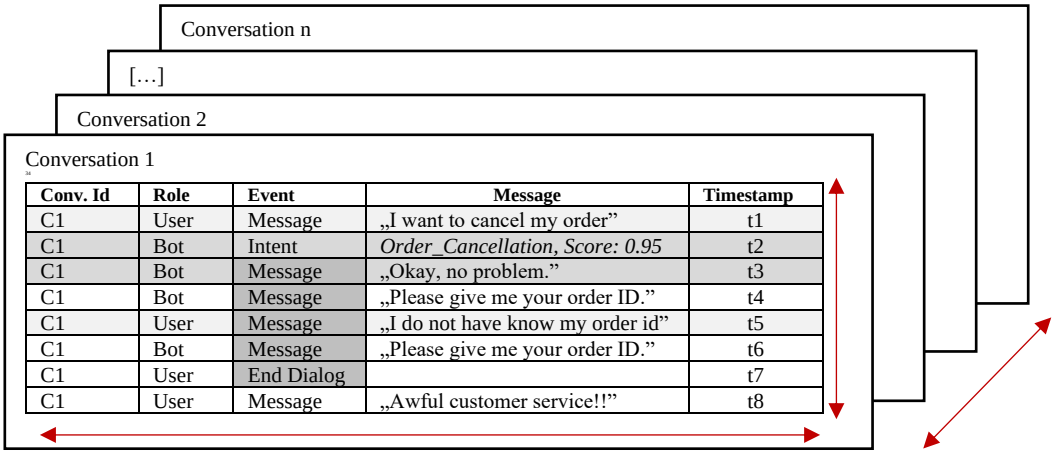
Conversational problems like these: concerns not understood, poor responses, or getting stuck in dialogue situations have a negative impact on the perception of the customer service as well as the entire technology, which ultimately reduces long-term adoption and undermines the business case (Diederich et al., 2021; Riquel et al., 2021; Van der Goot et al, 2021). These problems can be avoided (e.g., by limiting the user options on the frontend; Ryu et al., 2020) or mitigated (e.g., through repair strategies; Ashktorab et al., 2019) to some extent, however, chatbot managers from customer service as well as developers adressing the technical aspects (e.g., NLU) must first be aware of typical problems and potential improvements. For this purpose, chatbot analytics provides an excellent source of information, as the analysis of usage data covers the objective chatbot performance in many real-world customer chatbot conversations, can be strongly automated and comes at a low cost (Beaver and Mueen, 2020).

2.2 Chatbot Analytics

Chatbot managers and developers can use insights from chatbot analytics to improve their chatbots and reduce potential customer frustration (Følstad and Taylor, 2021; Riquel et al., 2021). Since there is a variety of behaviors as well as error causes and indicators, it is important to analyze conversations on different levels (Akhtar et al., 2019, Beaver and Mueen, 2020; Li et al., 2019).

Figure 2 shows an example chatlog and the different levels of chatbot analytics (Li et al., 2019; Rozga, 2018). On the event level, a single event (one row in **Figure 2**) is analyzed, e.g. when looking at a single utterance (e.g., *"I want to cancel my order"*). At the turn level, an utterance is analyzed in combination with the classified intent (score) and the chatbot response (e.g. "Classified Intent: Order_Cancellation Score: 0.95" and "*No Problem. You can cancel your orders using this link"*; Beaver and Mueen, 2020; Stolcke et al., 2000). At this level, many studies have already done research on conversational problems, e.g. evaluating strategies in case of a chatbot breakdown caused by the NLU and a low intent recognition score (Benner et al., 2021). Beyond the utterance/turn level, events can also be aggregated and analysed troughout multiple conversations, e.g. when calculating the average number of utterances or average intent recognition scores across all conversations (Przegalinska et al., 2019). Yet, as chatbot analytics research has found, some errors are not directly indicated and detectable automatically (Kvale et al.,

2020), e.g. filtering subjectively poor responses or intent mismatches, i.e. when a false intent was classified and logged with a sufficiently high enough score (false positive; Folstad and Taylor, 2021).



Conv. Id	Role	Event	Message	Timestamp
C1	User	Message	„I want to cancel my order"	t1
C1	Bot	Intent	*Order_Cancellation, Score: 0.95*	t2
C1	Bot	Message	„Okay, no problem."	t3
C1	Bot	Message	„Please give me your order ID."	t4
C1	User	Message	„I do not have know my order id"	t5
C1	Bot	Message	„Please give me your order ID."	t6
C1	User	End Dialog		t7
C1	User	Message	„Awful customer service!!"	t8

Figure 2. Chatlog illustrating the single and multi-event and conversation level (Li et al., 2019)

However, chatbot analytics research has pointed out that problems (i.e. opportunities for improvement) often occur in the progress of conversations (Beaver and Mueen, 2020). For example, customers repeat their words or trigger an intent several times if they are not understood, they abort the conversation or subdialogs if they get stuck, or they send (incorrect) inputs or user forms several times in a row, when they are stuck in a loop (Freeman and Beaver, 2017; Schloß et al., 2022). This cannot be determined in the chat log data at the single-event level, such as a single utterance (row) or by comparing the event over many conversations, instead the course of the conversation must be examined "vertically" (Takeuchi et al., 2007; Yaeli and Zeltyn, 2021). To date, however, there has been a lack of research in chatbot analytics on methods for analyzing conversational flows/processes, especially in easy-to-digest and visual form (Yaeli and Zeltyn, 2021). In particular, the analysis of different sequential events with attributes over many conversations, as illustrated by the 3 arrows in **Figure 2**, remains a challenge (Beaver and Mueen, 2020; Yaeli and Zeltyn, 2021). These process-oriented analyses, though, carry the potential to easily reveal the already mentioned problematic patterns (e.g., loops, aborts), to make conversations comparable for specific events that occur (e.g., handovers to human service employees), or to generally visualize conversation and topic flow (Freeman and Beaver; Schloß and Gnewuch, 2022). Additionally, a process-oriented monitoring and assessment of the „work" of chatbots, as it is already know from human service agents, is needed since many customer service chatbots operate at a high degree of autonomy, aiming to cover business processes end-to-end via interfaces to other systems (De Keyser et al., 2019; Meyer von Wolff et al., 2019; Takeuchi et al., 2007).

2.3 Process Mining

The analysis of sequential event data, such as that generated also in conversations, is the origin of process mining. Process mining generally is a technique to reconstruct and analyze business processes based on log data (Van der Aalst et al., 2011). The discipline located at the intersection of business process management and data mining is based on systems that generate large amounts of (event) log data, e.g. ERP systems or workflow management systems. The goal of process mining is to improve operational performance by exposing and analyzing existing processes and procedures (Song et al., 2008). This may involve, for example, the exploration and discovery of unknown processes or the comparison of existing process models with reality (i.e. real logged event data) (Topol, 2019).

Accordingly, there are 3 basic types of process mining: 1. Process discovery 2. Conformance checking and 3. Process enhancement (Van der Aalst et al., 2012). During process discovery, processes are reconstructed from log data without assumptions or models about the processes. For conformance checking, a process model already exists and the log data is checked for conformity. In the case of

process enhancement, the process model is not only compared to reality, but also extended by it. Ideally, this results in a better model of a real process (Van der Aalst et al., 2016).

The necessary prerequisite for process mining is the transformation of unstructured raw data into an event log (table) (Van der Aalst et al., 2012). This event log contains individual events on the row level, which necessarily need to be described by the the three attributes/columns case identificator, activity and timestamp. 1. First, the case identificator (case ID, e.g. „conversationId" in **Figure 2**) is needed to distinguish individual cases that will be aggregated later. 2. Second, defined activities are required, i.e. discrete steps forming processes that can be analyzed over the cases. 3. Third, a timestamp is needed to construct the sequence of events and activities. This can be a timestamp logged in a standard timestamp format (e.g. „2020-02-28T13:08:04.571Z") or a position identifier (e.g. „40439"), which would neglect the time difference or duration between several activities. In addition, further attributes in the event log can be used to extend the selection/filter functions or the description of the data or activities.

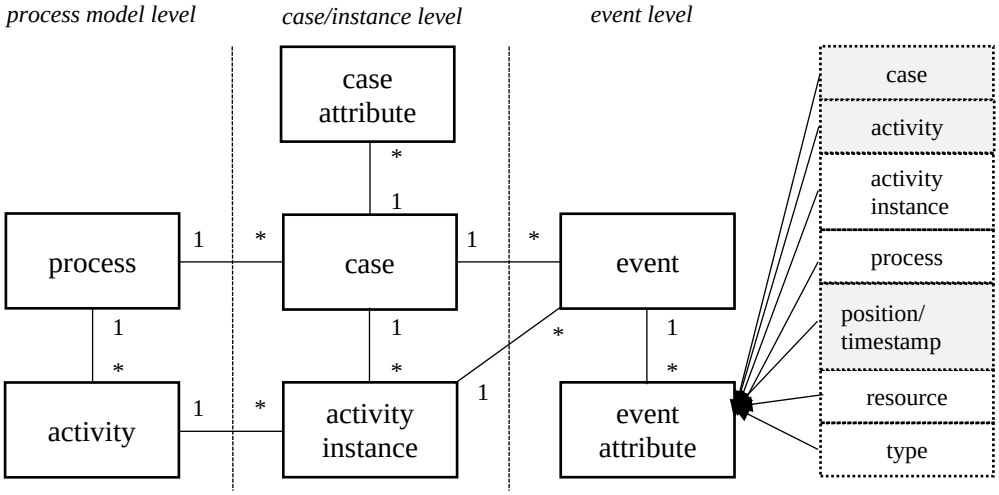


Figure 3. Class Diagramm of Process Mining (Van der Aalst, 2016)

Figure 3 shows the relationship of the process mining data in the form of a class diagram (Van der Aalst et al., 2016). At the process level (left), several activities (of a unique sequence) form a process; this process can be found in several cases (Van der Aalst et al., 2012). On the case level (middle), a case is composed of different case-individual activities and also has case attributes that describe the case as a whole (Sim et al., 2021). On the event level (right) an activity can consist of several smaller events, which are described by different attributes. To use process mining methods, some of the event attributes definitely must be logged to construct the event log, others are additional (Van der Aalst et al., 2011). Mandatory is the position or the timestamp of the events to construct the sequences as well as a case identifier for each event („event15 appeared in case4"). Additionally, the events need to be classified (summarized) into specific activities („event15 is part of activity2"). If these conditions are met and the data quality is sufficient, process mining algorithms generate calculations and visualizations (Berti et al, 2019; Topol, 2019). These then show the occurrence of processes, formed from unique sequences of activities (consisting of events), across a set of cases (Van der Aalst et al., 2011). In the following chapters, we explain how we applied this cardinalities to chatbot conversations to generate conversation- and process-oriented insights via conversation mining.

3 Method: Design Science Research

The conversation mining research project follows the design science research (DSR) approach. DSR ensures that the development process of a proposed solution is both scientifically rigorous and practically relevant by repeatedly collecting and incorporating real user feedback during the course of a research project (Gregor and Jones, 2007; Hevner et al., 2004). Accordingly, our DSR project is conducted in collaboration with an industry partner. The industry partner is a medium-sized service

provider for the energy industry that offers consulting, marketing or expert services to energy companies and develops digital solutions such as cus-tomer service chatbots. Through our industry partner, we had access to internal experts (chatbot developers), chatlog data, as well as their B2B utility customers, (customer service) chatbot managers. The chat data consisted of over 100,000 real-world customer chatbot conversations. We used these opportunities for practical insights especially when framing the problem, designing a solution, and evaluating and approving it.

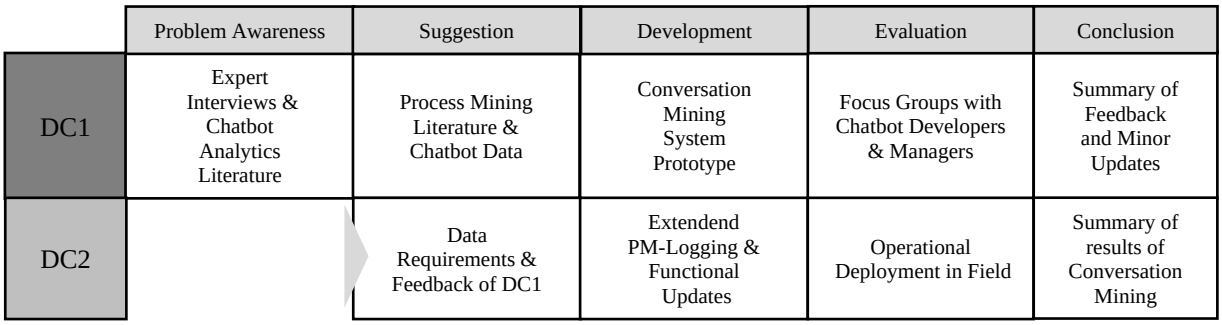


Figure 4. Structure of the Design Science Research Project (Kuechler and Vaishnavi, 2008)

Our DSR project is divided into two design cycles with five phases (**Figure 4**; Kuechler and Vaishnavi, 2008). This paper reports on the first design cycle of our DSR project, which was dedicated to the development of a first conversation mining prototype as an artifact and its evaluation. Starting with the problem awareness phase, we interviewed six chatbot experts regarding the challenges of customer service chatbots as well as problems and requirements for the analysis of critical or interesting conversations. We also reviewed the chatbot analytics literature to identify the current state of research as well as potential gaps. In the next phase, we suggested the use of process mining techniques to address the gaps in automated analysis at the conversational and process levels. Our actionable suggestions and steps were informed by the process mining literature and also inspired by the available chatbot log data from over 100,000 conversations of over 30 customer service chatbots deployed in energy industry. Subsequently, we developed an initial prototype to illustrate and instantiate our novel approach. The conversation mining system prototype, presented in the following chapters, was also evaluated during the 4th phase of our DSR project. For the evaluation, we collected feedback from 13 chatbot developers and managers in four focus groups. Since the initial feedback on our conversation mining system prototype was positive and encouraging, but also revealed areas of potential improvement, we plan to further refine the conversation mining system in a second design cycle. On the one hand, the benefits of the system can be further increased if the raw data logging, for which we already defined requirements, is expanded and optimized for process mining. On the other hand, the chatbot experts mentioned several functions and features to improve the system, which we will add (e.g., an export function). We are also planning to deploy our conversation mining system in DC2 so that it can be used and evaluated on an operative basis. Ultimately, the conversation mining system aims to close the loop between operational processes (conversations) and (chatbot) analytics, allowing users to extract data insights that can be used to improve the chatbot and its features (Quafari and van der Aalst, 2012). Moreover, we aim to gain and provide justificatory design knowledge from and for practical implementation (Gregor and Jones, 2007).

4 Designing a Conversation Mining System

4.1 Problem Awareness

To gain an initial understanding of the challenges that chatbot developers and managers face in designing and improving customer service chatbots, we conducted six one-hour one-on-one interviews with chatbot experts. The interviews were recorded and transcribed via Microsoft Teams and, as a follow-up, the state-ments were coded and finally matched to categories (chatbot performance causes and indicators, analytical challenges and requirements) in a closed coding process (Corbin and Strauss,

1990; Corbin and Strauss, 1998). The experts were four chatbot developers of the industry partner (in the roles of product owner, operations manager, UI/UX expert and NLU engineer) and two chatbot managers (in the roles of digitization and systems manager and executive director) of an energy provider deploying the customer service chatbot of the industry partner. Regarding chatbot performance, chatbot developers and managers consistently stated that, according to Chapter 2, a successful conversation consists of identifying a customer's concern and either providing an appropriate response or completing processes that are started in the chatbot. Conversa-tion failure, on the other hand, can have many causes, from NLU issues to stagnant dialog flow to inappro-priate responses or technical issues, which they all want to observe and track (Følstad and Taylor, 2021). The chatbot developers told us that they already used manual analysis, such as investigating unrecognized utterances (due to low intent scores). What was more difficult, however, due to inadequate logging, concepts, and evaluation methods, were specific insights into sources of error in conversation flows. For example, the operations manager and product owner emphasized that while performance was approximated with a completion rate, this rate only relied on static start and end points. This was also confirmed by looking at the data and the current chatbot analytics method: The status quo of the reporting used by the chatbot developers, which was also provided to the chatbot managers, consisted of comparing the number of final "success" events achieved (e.g., a form submitted) versus the start events of the business processes repre-sented in the chatbot dialogs. This approximation left only an interpretive ratio that did not provide insight into the actual problems and paths of customers. This gap was also evident in the requests of the chatbot managers who want to optimize the chatbot for customer service: For one, they wished to better understand which concerns (intents) the chatbot could handle autonomously and which not. In addition, the chatbot managers were very interested in being able to track the conversations that ultimately resulted in a handover to a human live chat agent. These requests indicated that the biggest analytical problem was connecting individual events to the overall context of the conversation and other events and their attributes. An example of this would be: out of 120 conversations that were handed over to live chat, 70 conversations included "questions about the invoice" (intent) the chatbot could not answer. Regarding chatbot performance in sub-dialogs or processes, both chatbot developers and managers wished for a more detailed breakdown of the reasons for errors, analogous to a funnel principle, for example: out of 100 people who started the rate calculation process, 20 ended the dialog directly, 15 did not enter their zip code correctly, and 65 were successfully advised. Furthermore, the chatbot developers mentioned loops and erroneous dialog sections as reasons for errors they would want to explore.

Common to all of the problems and requirements mentioned in the interviews was that although event-based data from conversations was logged en masse, chatbot developers and managers found it difficult to quickly gain insights into the conversational processes at scale. In contrast, the existing and prevailing approaches were a) aggregating simple metrics (e.g., number of messages) b) manually analyzing chat logs to gain an overview or c) defining static events as start and target values for success indications. With this in mind, we realized that a flexible approach was needed to allow to analyse the sequences of flexible events over many conversations – in the best case interactively. This need for automated approaches to conversation analysis beyond turn-level (Beaver and Mueen, 2020; Freeman and Beaver, 2017) and at best in graphical/visual form (Akthar et al., 2019, Yaeli and Zeltyn, 2021) has also been strongly emphasized in chatbot analytics research. Therefore, we came up with the flexible process-oriented method of process mining to inform the conversation mining project, to leverage data and address the shortcomings of current analytical approaches.

4.2 Suggestion

In order to develop our artifact, we followed an established procedure from process mining as depicted in **Figure 5**. We build on the process mining L* life-cycle as the justificatory knowledge for our design (Gregor and Jones, 2007; Van der Aalst et al., 2011). Starting with the justification of a project in Stage 0, the process mining L* life-cycle involves extracting all relevant data for *business and data understanding* required for the execution of the process mining project (Stage 1). This can be historical log data, open questions or already constructed KPIs and models. After we had found strong justification for the process mining approach in the problem description phase, we first worked out the *business*

understanding in Stage 1 (Van der Aalst et al., 2012). Therefore, we mapped the process mining goals of „process discovery", „conformance checking" and „model enhancement", outlined in chapter 2.3, to the design requirements identified from chatbot analytics research and practicioner feedback:

(1) The first major functionality of the proposed conversation mining system is the automated review of multiple conversations paths, corresponding to process discovery. Thus, the objective is not to confirm assumptions, but to explore and obtain an overview. Regarding chatbots, we propose that users of the system should be able to explore which topics (intents) customers often mention within a single conversation or in sequence (e.g., „Do customers ask for an invoice after asking about price increases?").

(2) The second major functionality for conversation mining is the comparison of ideal paths and reality according to conformance checking. Here, the focus should be more on assessing the performance of the chatbot. This is primarily applicable to guided dialogs, i.e. funnel-like dialogs with the chatbot, through which customers also conclude real-world business processes. For example, a customer updates contact data by several authentication and input steps in a chatbot dialog. As an ideal process model, we defined allowed sequences of activities that are the shortest paths to completion of a process/dialog. A deviation/distance indicates that customers had problems and a dialog, designed based on assumptions of the chatbot developers or managers, may not be optimally designed, for example in terms of guidance (e.g., how the messages are formulated) or the amount of dialog steps (Song et al., 2008).

(3) The ultimate goal of conversation mining based on process discovery and conformance checking, as indicated in the previous steps, should be chatbot optimization through well-informed chatbot developers and managers. Enhancement, in the context of conversation mining, therefore aims not only to the extension of the model (or the logs), but also to actual design improvements for the real process, dialog or conversation experience.

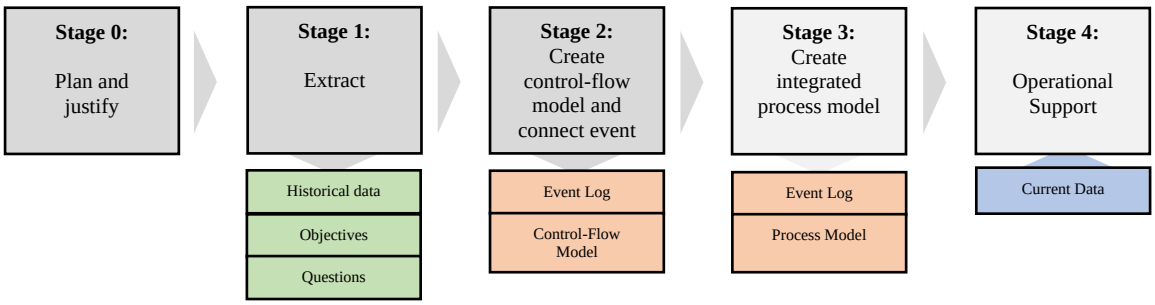


Figure 5. Five stages of the process mining L life-cycle (Van der Aalst et al., 2011)*

Once the idea of the project was outlined, in Stage 1 we turned our attention to the log data structure and quality *(data understanding).* First, we mapped the available conversational data to the general components of the process mining event log (introduced in the class diagram in Fig. 2, chapter 2.3) so that we could generate the analyses we had previously stated as a goal (Sim et al., 2021).

According to **Table 1**, we chose the conversations, identified by unique conversation IDs, as cases for both approaches. In addition, we found that all events (e.g., bot responses) were sufficiently logged in a timestamp format. Subsequently, one of the biggest initial design questions for the prototype was which activities could and should be evaluated at the process level (Van der Aalst, 2012). We decided to explore different activity levels. On the Process Discovery side, we chose the sequence of topics (intents), dialog steps of guided dialogs, and guided dialogs themselves as possible activities (Topol, 2019) . For Conformance Checking, we focused only on dialog steps (in particular: ideal transitions (Song et al., 2008)) comparing ideal (shortest) paths of a guided customer-chatbot dialog with the real-world interactions. As we found, the selection of the activities correlates strongly with the degree of variability. For example, in guided dialogs, customers can be restricted in their „degrees of freedom" because they can only cancel or continue the process. In open situations in conversations, on the other hand, customers have a variety of interaction options with the chatbot. If the intent is selected as the activity, customers can generate a variant number equal to the number of intents with their very first action („ask anything"). Theoretically, the number of variants is not limited since a process variant is as

a unique sequence of activities, e.g. 3 variants: A-B, A-B-C, A-C-B (Van der Aalst, 2011). With respect to data quality, in stage 1 and during the creation of the event log and the control flow model in stage 2, we made the following observations known from the process mining literature: First, each event meant to be associated with an activity must be logged with a reference. Since identifiers for dialog positions were not yet logged in the raw data, they were reconstructed for the prototype. In order to use the chatbot responses as a proxy for progress in a guided dialog (customer arrives at step n), we matched the chatbot responses logged as text with configuration database entries (see *, Table 1) to determine the affiliation to dialogs or dialog steps. Second, by creating a control flow diagram, we determined which events or attributes were still insufficiently logged. We drew the complete process of a representative and particularly complex guided dialog with all distinct real-world options a customer has on the interface. We found there were some differences between the real-world events and logged events, which process mining research recommends closing. If logging is insufficient, the logged variants do not reflect the real variance in customers' behavior (interactions). Examples are link clicks logged at the conversation but not event level or a missing identificatior for different input types as an event attribute. For this reason, we defined the requirements for an enhanced logging informed by the event log best practices. In Design Cycle 2, we plan to add an enriched dataset as recommended for Stage 3 of the L* life-cycle.

	Process Discovery	Conformance Checking
cases	conversations (conversationId)	conversations (conversationId)
timestamp	timestamp	timestamp
activity	topic, dialog step*, dialog*	dialog step*
process	sequence of activities	sequence of activities
events	intent, bot responses (* and configs)	bot responses (* and configs)
event attributes	type, customer	type, customer
variance	high	low

Table 1. Application of Process Mining Classes on Conversational Data

4.3 Development of the Conversation Mining System

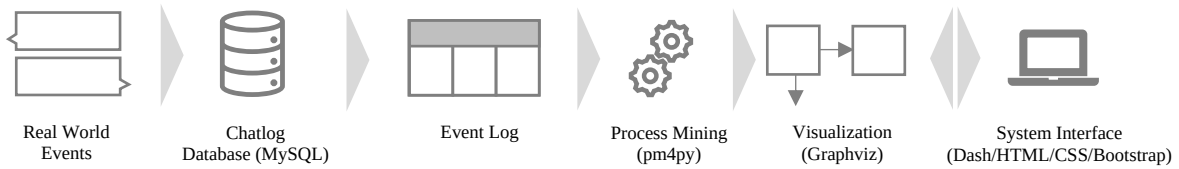


Figure 6. Architecture of the Conversation Mining System

The starting point for the conversation mining system are the real world events in the chatbot conversations, which should be reflected as accurately as possible in the log data according to the requirements of process mining. As depicted in **Figure 6**, our data base consists of two MySQL data-base tables, one on the conversation level, one on the event level with more than five million events. A MySQL query selects the relevant information from the database. The query transforms the raw data to a format where each chatbot message is mapped to a corresponding intent, dialog or dialog step. In the next step, the event log is created, consisting of caseID, activity key and timestamp (as well as further attributes). For the process mining, we use the open-source python library pm4py to calculate/generate the statistics, processes and variants and Graphviz to visualize them (Berti et al., 2019). Ultimately, these process visualizations are embedded in a multipage dashboard application. Using callbacks, it is possible for users, i.e. chatbot developers and managers, to dynamically vary parameters on the interface and reload the visualizations over and over again. Thus, and through the structure of the prepared data, the selected activity key can also be varied. Users are provided with the possibility to adjust the granularity of the analysis as desired which expands their exploration and interpretation possibilities. The process discovery page, shown on the top in **Figure 7**, has different functionalities to support the user to comfortably explore conversations and their paths. He or she can interact with several interface

ele-ments such as dropdowns, radio items, or sliders. The interface elements on the discovery page are organized into three main groups: 1) a left column with filters; 2) a center column for the visualiza-tion of the conversation flow processes, and 3) a right column with counters and two charts. On the left, the date range and the customers (different chatbots) can be filtered. As mentioned above, the activity level (intent, dialog, dialog step) can be adjusted, and, depending on the use case, activity filters (starts, ends or contains activity) can also be applied. In addition, the process mining algorithm and thus graph can also be selected (Heuristics Miner, Directly Follows Graph, Process Tree), with the Directly Follows Graph (DFG) providing a good overview for the case of conversation mining. The process graph gets updated dynamically, also with multiple filters. Starting from the first node of the graph, the numbers on the arrows show the number of transitions between two activities summed over all conversations, the numbers in the color coded rectangles represent the frequencies of the respective activities. On the right, the number of conversations and the number of variants currently displayed are also refreshed dynamically. Below, there is a variant bar chart, representing the distribution of process variants among the conversations, combined with a variant slider on the right. As with established process mining tools, users can choose the number of variants they want to select and see. This is particularly important for high numbers of possible variants, as transparency could suffer if all variants were displayed („spaghetti process", Van der Aalst, 2012). Additionally, a pie chart depicts the shares of the last activities:

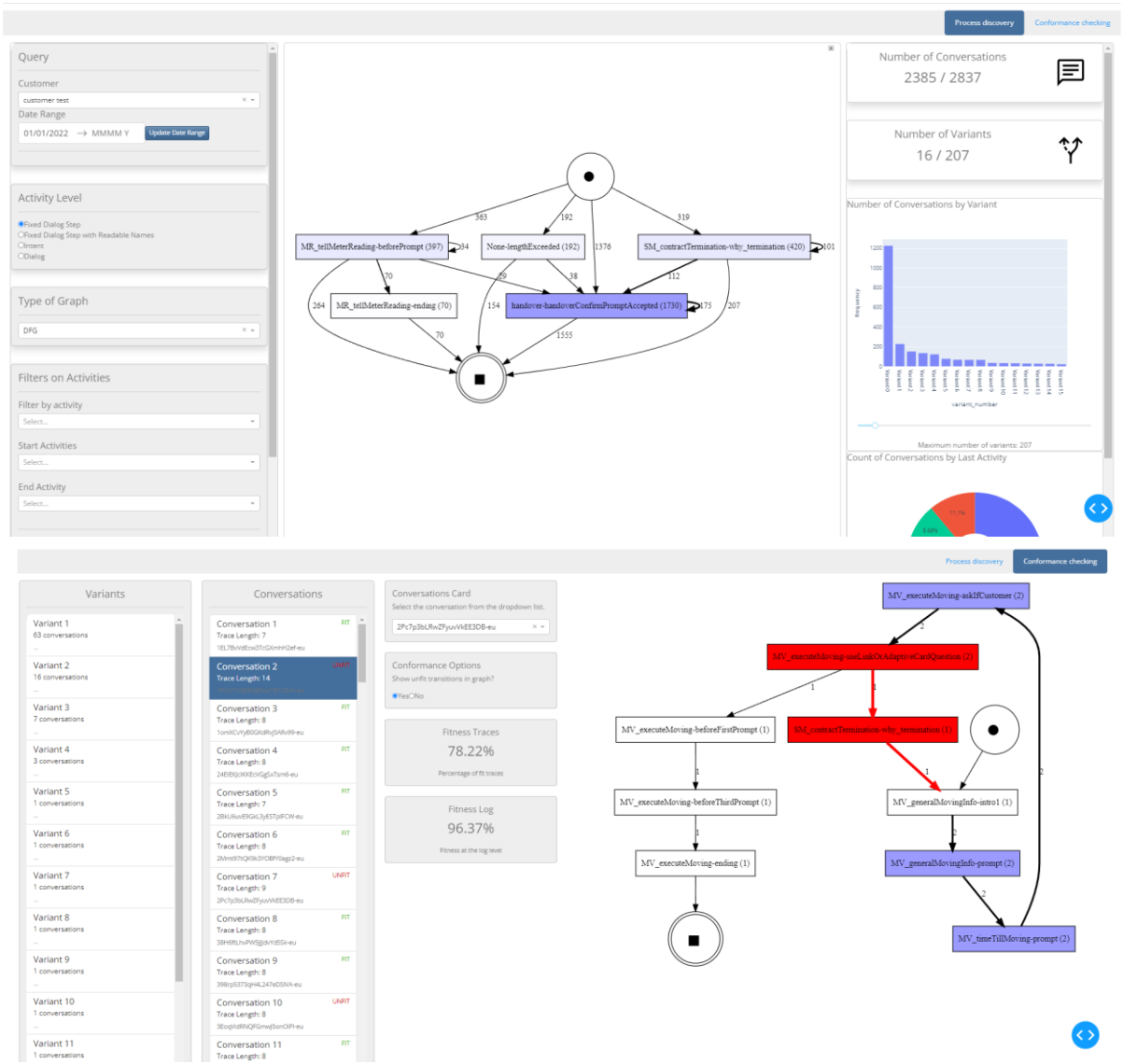


Figure 7. Process Discovery and Conformance Checking in the Conversation Mining System

On the conformance checking page, shown at the bottom half of **Figure 7**, users can compare the performance of guided dialogs with an ideal process model. The activities of the optimal dialog are predefined. The page contains two columns (variants and conversations), the conversation graph, and fitness diagnostic stats. The variants column groups all the variants for the current log and presents them in individual cards. For each variant, a specific number of conversations has an equal process flow. The conversation column groups all the conversations of the selected log and each conversation card provides the values conversation id, trace length, and conversation fitness. The trace length is the sum of transitions between dialog steps (activities) (Song et al., 2008). A trace is „fit" if an activity is followed by one provided in the ideal model, a conversation is „fit" if only contains „fit" traces (Song et al., 2008). Since a dialog may have several ideal paths, a higher trace length does not necessarily imply unfitness, because a customer could have chosen a subprocess that simply consists of more steps by default (Berti et al., 2019). The user can select conversations to inspect them by clicking on the corresponding card or via the dropdown. To allow visual highlighting of unfit traces directly on the graph, the conformance functionalities of pm4py were extended by accessing the dot language source code of the pm4py-generated graphs (Berti et al., 2019). Additional functions also modify the properties of nodes and edges (e.g. background color, edge thickness) for visualization.

4.4 Evaluation and Results

To evaluate our conversation mining prototype and the underlying analytical approach, we conducted four focus groups with 7 chatbot developers and 6 chatbot managers in total. **Table 2** provides an overview on our participants. A focus group is qualitative social research method and is used to obtain artifact feedback from a focused conversation with a largely homogeneous group (with individual perspectives). Focus groups are designed to encourage participants to express thoughts and feedback in a structured process (Morgan, 1996). Our focus groups lasted from 1 to 1.5 hours and consisted of a 10-minute introduction, a 10-minute demonstration of the prototype, an open Q&A session with initial verbal feedback and a SWOT analysis for explicit feedback during the last 20 minutes. We chose the SWOT analysis since it is particularly useful for recording individual perceptions because it contains different perspectives of feedback (strengths, weak-nesses, opportunities, threats) and encourages participants to reflect on all of them (Helms and Nixon, 2010). In addition, writing down the points individually helps avoiding participant's opinions being unheard or overruled, which can happen with a purely verbal survey (i.e., groupthink). To ensure equal opportunity for everyone to give feedback, we took care to avoid the individual focus groups being too large, as recommended by best practices (Morgan, 1996). The resulting 4 focus groups were conducted either physically (chatbot developers) or remotely (chatbot managers) and were recorded and transcribed using MS Teams. The transcripts as well as the SWOT notes were coded by two authors first individually, then jointly (open to axial coding) and finally assigned to categories (Corbin and Strauss, 1990; Corbin and Strauss, 1998).

Group	Role	Job Role
G1	Chatbot Developers	Software Engineer (EX1), Software Engineer (EX2), DevOps Engineer (EX3)
G2	Chatbot Managers	Online Marketing Manager (EX4), Online Marketing Manager (EX5), Head of Marketing (EX6), Online Marketing Manager (EX7)
G3	Chatbot Developers	Head of BPO & Software (EX8), Product Owner (EX9), Customer and Operations Support Manager (EX10), Natural Language Processing Expert (EX11)
G4	Chatbot Managers	Digitization and Systems Manager (EX12), Executive Director (EX13)

Table 2. Participants of the four focus groups

The feedback we received on the conversation mining system prototype was overall very positive, but also contained suggestions regarding potential for further development. Due to the different roles of the experts, we received insights from very different perspectives. In line with the suggestions in the chatbot analytics literature (e.g. Akhtar et al., 2016; Yaeli and Zeltyn, 2021), many focus group participants highlighted the visual approach as particularly positive (as can be seen in the results of the SWOT analysis in **Table 3**). The conversation mining system was evaluated as helpful to analyze large amounts of data (EX6) and to reveal conversations paths. It may also support new employees in understanding

dialog structures (EX11). It was also emphasized that the system itself as well as findings require some knowledge and interpretation (EX8, EX6). Nevertheless, the experts found and mentioned many practical use cases, such as the simplified identification of dialog aborts, the investigation of handover situations (EX13) or the application on dialogs with high error rates to recognize reasons for loops and aborts (EX9) (Quafari and van der Aalst, 2012). Many participants particularly emphasized the potential for operational decision support. They reflected on concrete improvements for the chatbot design, for example, the deletion of unused or unnecessary intents or steps or potential improvements for new found bottlenecks in often uncompleted dialogs. For example, it became apparent that some of multiple dialog paths were never selected, while others were associated with high dropout rates (EX8). Especially in the long term, the experts attributed positive effects to the conver-sation mining approach, also because product management is equipped to make better (more informed) decisions (EX1). The software developers also commented on future effort for appropiate logging before the full potential of the approach is realized (EX1, EX2). Still, they acknowledged that this analysis can also be an incentive to further modularize and improve the data models, data structure and ultimately components of the chatbot (EX13), e.g. regarding the logging of error/validation messages within dialogs. As the process mining literature and these statements suggest, conversation mining is not just an arbitrary analytics layer. Instead, it promotes to reconsider data preparation and modelling and can initiate a continuous improvement process through an integrated model-reality feedback loops (Van der Aalst, 2012).

Strengths	*Weaknesses*
Raw Data Logging: - leveraging more/all of the available data **Event Log Construction:** - Degree of analysis (rough to fine analysis) can be determined independently - Can be applied to extensive amounts of data **Deployment:** - Prototype can be in use quickly **Usability/Interface:** - Professional, intuitive and clear layout of the prototype (2) - Good and clear visualizations for analysis (5) - Very good approach/tool for analyzing conversational steps/ processes (2) **Interpretation** - Good recognition of abortions/abortion paths - Connection between different sub-dialogs is appearant - Improvement of the user experience - Detailed evaluation and optimization options for users	**Raw Data Logging:** - Increased effort when developing dialog steps - Changes in initial logging cannot be applied to past historical conversations **Event Log Construction:** - User utterances during guided dialogs that were not answered should also be integrated - Needs to be filtered per customer/chatbot (2) - Accurate aggregation/process modell of events **Usability/Interface:** - Non-Technical Names for Intents/Dialogs needed - Clarity is lost as soon as variance is high - Minor technical limitations **Interpretation** - Some know-how needed for drawing conclusions (2)
Opportunities	*Threads*
Raw Data Logging: - Improved quality of technical requirements for logging and new dialogs (positive feedback loop) (2) - Better data preparation for customer reportings **Event Log Construction:** - Include clicked links as an activity - Allow filtering for visitor-URLs **Deployment:** - Integration into existing analytics tool (PowerBI) - Create interfaces to databases and other tools **Usability/Interface:** - User friendliness, overall overview - Export Function (2) - Directly connect view of the actual conversations **Interpretation** - Identification of causes of poor success rates - Identify utilization of dialogs (2) - Obsolete dialog steps can be discussed and altered (2) - Processes/dialogs can be adapted based on results (3) - Improved quality and evidence-based definition of technical requirements of the product management (2) - Additional use cases: marketing purposes, analyzing other conversational data (2)	**Raw Data Logging:** - high effort for robustness/correctness of logging (2) **Event Log Construction:** - Current state relies on matching between logged answers and manually maintained database entries **Deployment:** - Too many tools in general might overwhelm **Usability/Interface:** - Speed/refresh times when loading data - Data Security Concerns (private data shown) **Interpretation** - Results and interpretation depend on the chatbot software and its configuration - Could be overwhelming for non-tech chatbot managers

Table 3. Results of the SWOT analysis of all focus groups

5 Discussion and Conclusion

In this paper, we presented a DSR project dedicated to the design of a conversation mining system to support chatbot developers and managers in improving their customer service chatbots via visual chatbot analytics at the process level. We first mapped out the problem space through expert interviews and related research, after which we used process mining methods and data available to us to conceptually design the conversation mining. Subsequently, we developed the associated conversation mining system as a prototype and conducted four focus groups with customer service and chatbot experts. The results of the evaluation suggest that our method, as well as its instantiation in the artifact, have great potential to provide chatbot developers and managers with novel in-depth insights into conversational processes and provide them with findings for possible improvements in operative chatbot management.

5.1 Theoretical and Practical Contributions

First, our research contributes to the literature on chatbot design and analytics by providing an innovative approach for analyzing chatbot data to gain insights for chatbot improvement. Specifically, we provide design knowledge for conversation mining approaches and systems that enable process-level analysis of conversation data (Beck et al., 2013; Gregor and Jones, 2007). This extends existing methods such as interviews or manual or single event-based chatbot analytics by demonstrating a more scalable approach to gaining detailed insights into the performance of a customer service chatbot (Følstad and Taylor, 2021; Przegalinska et al., 2019; Yaeli and Zeltyn, 2021). Second, we offer interesting insights into the applicability and utility of using established process mining methods in the context of chatbot conversations. Overall, we find that such an approach can help to better visualize the complex nature of conversations also named by chatbot research (Schuetzler et al., 2021; Yaeli and Zeltyn, 2021). This also extends existing research that has used process mining on various types of (text-based) data (e.g., Holstrup et al., 2020; Kecht et al., 2021; Topol, 2019). Specifically, we reveal insights on how event logs need to be created for deploying process mining for custchatbots. Finally, we shed light on the challenges of extracting insights from chatbot data for chatbot developers and managers. They often struggle to derive improvement measures and make specific informed decisions based on the chatbot operation as the actual chatbot performance and its drivers remain hidden (Kvale et al., 2019). Our results show that our tool can help them. On the practical side, we contribute the system itself, as well as use cases to leverage it. On the one hand, conversation mining serves for conversation exploration. The activities can be selected variably; thus, for example, it is possible to examine which intents/topics customers access one after the other in which frequency (Topol, 2019). With appropriate event log data, interaction types can also be examined, for example, button clicks vs. text inputs. In addition, the system provides a variety of filters to detect the most (frequent) problematic conversational situations, via specific events, aborts, last activities, or loops (Beaver and Mueen, 2020). By means of our conformance checking we also provide an approach to define ideal processes of dialogs and to compare them with real processes on fitness. This makes the most problematic bottlenecks in the chatbot easily identifiable and helps chatbot developers and managers to adjust their chatbot, e.g. regarding the dialog flow.

5.2 Limitations and Future Research

Our research is not without limitations. Since the conversation mining system as an analytical layer is independent of the specific activities, its insight value is only as good as the underlying raw data (Van der Aalst, 2016). Since the data set available to us was not perfect (e.g., missing identifiers for dialog steps), we will address data quality and selection for the event log in DC2. Another limitation we address in DC2 is the qualitative evaluation (focus groups), which we plan to supplement with a quantitative case study analysis once the system is embedded and the event log quality further improved (Van der Aalst et al., 2007). Limitations also include the specific data and context. We therefore encourage chatbot analytics research to apply and test conversation mining in other industries and on other data. In particular, conversational agent interactions with limited degrees of freedom and clear processes could benefit from our method (Van der Aalst, 2011). Hence, voice bots are also a suitable subject for conversation mining in the future, since they follow multi-step predefined turns (Zierau et al., 2022).

References

Ashktorab, Z., Jain, M., Liao, Q. V., and Weisz, J. D. (2019, May). Resilient chatbots: Repair strategy preferences for conversational breakdowns. In Proceedings of the 2019 CHI conference on human factors in computing systems (pp. 1-12).

Akhtar, M., Neidhardt, J., and Werthner, H. The potential of chatbots: Analysis of chatbot conversations, Proceedings - 21st IEEE Conference on Business Informatics, 1, pp. 397–404. (2019)

Beaver, I. and Mueen, A. (2020). „Automated Conversation Review to Surface Virtual Assistant Misunderstandings: Reducing Cost and Increasing Privacy," *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (8), 13140-13147.

Beck, R., Weber, S., and Gregory, R. W. (2013). „Theory-generating design science research," *Information Systems Frontiers* 15(4), 637-651.

Berti, Alessandro & van Zelst, Sebastiaan & Aalst, Wil. (2019). Process Mining for Python (PM4Py): Bridging the Gap Between Process- and Data Science.

Benner, D., Elshan, E., Schöbel, S., and Janson, A. What do you mean? A review on recov-ery strategies to overcome conversational breakdowns of conversational agents, ICIS 2021 Proceedings. (2021)

Bordoloi, S., Fitzsimmons J., and Fitzsimmons M. (2019) *Service Management: Operations, Strategy, Information Technology*, 9th Edition. London: McGraw-Hill.

Brandtzaeg, P. B., and Følstad, A. (2017). Why people use chatbots. In Internet Science: 4th International Conference, INSCI 2017, Thessaloniki, Greece, November 22-24, 2017, Proceedings 4 (pp. 377-392). Springer International Publishing.

Brandtzaeg, P. B. and Følstad, A. (2018). „Chatbots: changing user needs and motivations," *Interactions* 25 (5), 38-43.

Corbin, J. M. and Strauss, A. (1990). „Grounded theory research: Procedures, canons, and evaluative criteria," *Qualitative sociology* 13(1), 3-21.

Corbin, J. M. and Strauss, A. (1998). *Basics of qualitative research techniques*, Thousand oaks, CA: Sage publications.

Dale, R. (2016). „The return of the chatbots," *Natural Language Engineering* 22(5), 811-817.

De Keyser, A., Köcher, S., Alkire, L., Verbeeck, C., and Kandampully, J. (2019). „Frontline Service Technology infusion: conceptual archetypes and future research directions," *Journal of Service Man-agement* 30(1), 156-183.

Diederich, S., Lembcke, T. B., Brendel, A. B., & Kolbe, L. M. (2021). Understanding the impact that response failure has on how users perceive anthropomorphic conversational service agents: Insights from an online experiment. AIS Transactions on Human-Computer Interaction, 13(1), 82-103.

Følstad, A. and Skjuve, M. (2019). „Chatbots for Customer Service: User Experience and Motivation," *Proceedings of the 1st International Conference on Conversational User Interfaces* 1 (1), 1-9.

Følstad, A. and Taylor, C. (2021). „Investigating the user experience of customer service chatbot interaction: a framework for qualitative analysis of chatbot dialogues," *Quality and User Experience* 6 (1), 1-17.

Freeman, Cynthia, and Ian Beaver. „Online proactive escalation in multi-modal automated assistants." *The Thirtieth International Flairs Conference*. 2017.

Galitsky, B. (2019) *Developing enterprise chatbots: learning linguistic structures*. 1st Edition. Cham: Springer.

Gao, M., Liu, X., Xu, A., and Akkiraju, R. (2021). „Chatbot or Chat-Blocker: Predicting Chatbot Popularity before Deployment," *Designing Interactive Systems Conference.*

Gnewuch, U., Morana, S., Adam, M. T. P., and Maedche, A. (2022). Opposing Effects of Response Time in Human–Chatbot Interaction. *Business & Information Systems Engineering*, 1-39.

Gregor, S. and Hevner, A. R. (2013). „Positioning and presenting design science research for maximum impact," *MIS quarterly* 37 (2), 337-355.

Gregor, S., and Jones, D. (2007). The Anatomy of a Design Theory. Journal of the Association for Information Systems, 8(5), 312–335.

Grudin, J. and Jacques, R. (2019). „Chatbots, humbots, and the quest for artificial general intelligence". *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* 209, 1-11.

Helms, M. M., & Nixon, J. (2010). Exploring SWOT analysis–where are we now? A review of academic research from the last decade. Journal of strategy and management, 3(3), 215-251.

Hevner, A. R., March, S. T., Park, J., and Ram, S. (2004). „Design science in information systems research," *MIS quarterly* 28 (1), 75-105.

Hevner, A., vom Brocke, J., and Maedche, A. (2019) „Roles of digital innovation in design science research," *Business & Information Systems Engineering* 61, 3-8.

Holstrup, A., Starklit, L., & Burattin, A. (2020, July). Analysis of Information-Seeking Conversations with Process Mining. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

Huang, D. H., and Chueh, H. E. (2021). „Chatbot usage intention analysis: Veterinary consultation." Journal of Innovation & Knowledge, 6(3), 135-144.

Janarthanam, S. (2017). *Hands-on chatbots and conversational UI development: build chatbots and voice user interfaces with Chatfuel, Dialogflow, Microsoft Bot Framework, Twilio, and Alexa Skills*. 1st Edition. Birmingham: Packt Publishing Ltd.

Janssen, A., Grützner, L., and Breitner, M. H. (2021). „Why do Chatbots fail? A Critical Success Factors Analysis," *International Conference on Information Systems*, Austin, Texas.

Jenkins, M. C., Churchill, R., Cox, S., and Smith, D. (2007). „Analysis of user interaction with service oriented chatbot systems," *International Conference on Human-Computer Interaction*, Beijing, China.

Jiang, L., Wang, X., Chen, Q., and Min, Q. (2020) „User Switching Behavior: AI Chatbots or Human Agents?," *Pacific Asia Conference on Information Systems*.

Johannsen, F., Leist, S., Konadl, D., and Basche, M. (2018). „Comparison of commercial chatbot solutions for supporting customer interaction," *European Conference on Information Systems*, Ports-mouth, United Kingdom.

Juniper Research. (2022). *Chatbots: Market Forecasts, Sector Analysis & Competitor Leaderboard 2022-2026.* Chatbot Reports.

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing*, 2nd Edition, New Jersey: Pren-tice-Hall.

Kecht, C., Egger, A., Kratsch, W., & Röglinger, M. (2021, October). Event Log Construction from Customer Service Conversations Using Natural Language Inference. In *2021 3rd International Conference on Process Mining (ICPM)* (pp. 144-151). IEEE.

Kucherbaev, P., Bozzon, A., and Houben, G. J. (2018). „Human-aided bots," *IEEE Internet Computing*, 22 (6), 36-43.

Kuechler, B., and Vaishnavi, V. (2008). On theory development in design science research: anatomy of a research project. *European Journal of Information Systems, 17*(5), 489-504.

Kvale, K., Sell, O. A., Hodnebrog, S., and Følstad, A. Improving conversations: lessons learnt from manual analysis of chatbot dialogues, Lecture Notes in Computer Science, 11970 , pp. 187–200. (2020)

Li, C. H., Chen, K., and Chang, Y. J. (2019). „When there is no progress with a task-oriented chatbot: A conversation analysis," *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services* 59, 1-6.

McTear, M., Callejas, Z., and Griol, D. 2016. The Conversational Interface, Cham: Springer.

Meyer von Wolff, Raphael, Sebastian Hobert, and Matthias Schumann. „How may i help you?–state of the art and open research questions for chatbots at the digital workplace." Proceedings of the 52nd Hawaii international conference on system sciences. 2019.

Przegalinska, A., Ciechanowski, L., Stroz, A., Gloor, P., and Mazurek, G. In bot we trust: A new methodology of chatbot performance measures, Business Horizons (62:6), pp. 785 –797. (2019)

Quafari, M. S., and Van der Aalst, W. 2012. „Root Cause Analysis with Enriched Process Logs," in Business Process Management Workshops, M. La Rosa and P. Soffer (Eds.), Cham: Springer, pp. 174–186.

Riquel, J., Brendel, A. B., Hildebrandt, F., Greve, M., & Dennis, A. (2021). „F*** You!"–An Investigation of Humanness, Frustration, and Aggression in Conversational Agent Communication. ICIS 2021 Proceedings. 1.

Rozga, S. (2018). *Practical Bot Development - Designing and Building Bots with Node.js and Microsoft Bot Framework*, 1st Edition. Berkeley, CA: Apress.

Ryu, H., Kim, S., Kim, D., Han, S., Lee, K., & Kang, Y. (2020). Simple and steady interactions win the healthy mentality: Designing a chatbot service for the elderly. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1-25.

Schloß, D., and Gnewuch, U. (2022). Conversation Mining for Customer Service Chatbots. CONVERSATIONS 2022 Position Paper.

Schloß, D., Gnewuch, U., and Maedche, A. (2022). Towards Designing a Conversation Mining System for Customer Service Chatbots. ICIS 2022 Proceedings.

Schuetzler, R. M., Grimes, G. M., Giboney, J. S., and Rosser, H. K. (2021). „Deciding Whether and How to Deploy Chatbots", *MIS Quarterly Executive* 20 (1), 1-15.

Sim, S., Bae, H., & Liu, L. (2021). Bagging recurrent event imputation for repair of imperfect event log with missing categorical events. *IEEE Transactions on Services Computing*.

Song, M., Günther, C. W., & Van der Aalst, W. M. (2008, September). Trace clustering in process mining. In *International conference on business process management* (pp. 109-120). Springer, Berlin, Heidelberg.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., and Meteer, M. (2000). „Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech," *Computa-tional linguistics* 26 (3), 339-373.

Takeuchi, H., Subramaniam, L. V., Nasukawa, T., Roy, S., and Balakrishnan, S. 2007. „A conversation-mining system for gathering insights to improve agent productivity," Proceedings - The 9th IEEE International Conference on E-Commerce Technology; The 4th IEEE International Conference on Enterprise Computing, E-Commerce and E-Services, CEC/EEE 2007, pp. 465–468.

Topol, Zvi, Using Process Mining to Improve Conversational Interfaces", https://flux-icon.com/camp/2019/3, last accessed 2022/25/09.

Van Der Aalst, W. M., Reijers, H. A., Weijters, A. J., van Dongen, B. F., De Medeiros, A. A., Song, M., and Verbeek, H. M. W. (2007). Business process mining: An industrial application. *Information systems*, *32*(5), 713-732.

Van Der Aalst, W. et al. Process mining manifesto. International conference on business process management. Springer, Berlin, Heidelberg, (2011).

Van Der Aalst, W. Process mining: Overview and opportunities. ACM Transactions on Management Information Systems (TMIS), 3(2), 1-17. (2012).

Van Der Aalst, W. (2016). Process mining: data science in action (Vol. 2). Heidelberg: Springer.

Van der Goot, M. J., Hafkamp, L., and Dankfort, Z. (2021). Customer service chatbots: A qualitative interview study into the communication journey of customers. In Chatbot Research and Design: 4th International Workshop, CONVERSATIONS 2020, Virtual Event, November 23–24, 2020, Revised Selected Papers 4 (pp. 190-204). Springer International Publishing.

Yaeli, A., and Zeltyn, S. 2021. „Where and Why is My Bot Failing? A Visual Analytics Approach for Investigating Failures in Chatbot Conversation Flows," Proceedings - 2021 IEEE Visualization Conference - Short Papers, VIS 2021, pp. 141–145.

Zierau, N., Hildebrand, C., Bergner, A., Busquet, F., Schmitt, A., & Marco Leimeister, J. (2022). Voice bots on the frontline: Voice-based interfaces enhance flow-like consumer experiences & boost service outcomes. Journal of the Academy of Marketing Science, 1-20.