ICIS 2022 Proceedings

Service Science

Dec 12th, 12:00 AM

Towards Designing a Conversation Mining System for Customer Service Chatbots

Daniel Schloß
Karlsruhe Institute of Technology (KIT), daniel.schloss@kit.edu

Ulrich Gnewuch
Karlsruhe Institute of Technology (KIT), ulrich.gnewuch@kit.edu

Alexander Maedche
Karlsruhe Institute of Technology (KIT), alexander.maedche@kit.edu

Towards Designing a Conversation Mining System for Customer Service Chatbots

Short Paper

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

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

Alexander Maedche
Karlsruhe Institute of Technology
Karlsruhe, Germany
alexander.maedche@kit.edu

Abstract

Chatbots are increasingly used to provide customer service. However, despite technological advances, customer service chatbots frequently reach their limits in customer interactions. This is not immediately apparent to both chatbot operators (e.g., customer service managers) and chatbot developers because analyzing conversational data is difficult and labor-intensive. To address this problem, our ongoing design science research project aims to develop a conversation mining system for the automated analysis of customer-chatbot conversations. Based on the exploration of large dataset (N= 91,678 conversations) and six interviews with industry experts, we developed the backend of the system. Specifically, we identified and operationalized important criteria for evaluating conversations. Our next step will be the evaluation with industry experts. Ultimately, we aim to contribute to research and practice by providing design knowledge for conversation mining systems that leverage the treasure trove of data from customer-chatbot conversations to generate valuable insights for managers and developers.

Keywords: chatbot, customer service, conversation mining, design science research

Introduction

Thanks to advances in artificial intelligence, more and more companies are deploying chatbots in recent years (Dale, 2016; Schuetzler et al., 2021). Chatbots are software applications designed to interact with customers through text-based communication (Grudin and Jacques, 2019). Due to their natural language understanding (NLU) capabilities, they offer users a way to interact very naturally using written language (Hill et al., 2015; McTear et al., 2016). Chatbots are therefore often used to automate customer interactions and increasingly complement the frontline of customer service (Følstad and Skjuve, 2019; van Doorn et al., 2017; Wirtz et al., 2018). They offer 24-hour availability and can hold many conversations in parallel (Brandtzaeg and Følstad, 2017; Klopfenstein et al., 2017). As a technology with the potential for cost-oriented automation, chatbots are seen as a strategic element of present and future customer service (De Keyser et al., 2019; Thomaz et al., 2020).

However, despite the advances and potentials, the technology is being challenged in practice. On the one hand, chatbots are still limited in processing customer requests (Klopfenstein et al., 2017; Takayama et al., 2019). On the other hand, their actual performance in the course of a conversation is usually only observable in highly aggregated figures (Przegalinska et al., 2019). In the past, many long-term chatbot implementation projects have failed (Gao et al., 2021). To avoid this, both chatbot operators (e.g., customer service managers) and chatbot developers have a strong interest in gaining insights into the conversations and actual performance of an active customer service chatbot (e.g., in terms of successfully handled customer concerns). Customer service managers need insights into the chatbot's interactions with customers to assess service quality (Følstad and Taylor, 2021). Chatbot developers need to make informed technical improvements to the chatbot, such as retraining its language model (Cardoso et al., 2015), and review conversations for quality assurance (Beaver and Mueen, 2020). The most important source for these insights is the large amount of usage data from customer service chatbots (Følstad and Brandtzaeg, 2017). However, due to the relative novelty of chatbot technology, there is a lack of design knowledge for analytical systems that leverage data from customer-chatbot conversations in order to continuously improve the performance of the chatbot (Beaver and Mueen, 2020; Følstad and Taylor, 2021). For this reason, we conduct a design science research (DSR) project dedicated to the following research question:

RQ: How to design and develop a conversation mining system to assist chatbot operators and developers to continuously assess and improve the performance of customer service chatbots?

In our ongoing DSR project, we draw on the literature on human-chatbot interactions, especially in the context of customer service (e.g., Mozafari et al., 2022; Reinkemeier and Gnewuch, 2022). We specifically build on and extend prior chatbot analytics research (e.g., C. H. Li et al., 2020) to design and develop a conversation mining system that aims to close the gap between high-level metrics of customers' experience of an interaction, such as total number of messages exchanged (e.g. Przegalinska et al., 2019), or metrics that only refer to conversation parts, such as a single failure of the NLU (Benner et al., 2021). To this end, we account for the fact that customer-chatbot conversations follow processes where deviations between an expected or optimal and actual course can be analyzed (Takeuchi et al., 2007; Van Der Aalst, 2012; Yaeli and Zeltyn, 2021). To design a conversation mining system that fits this purpose, we analyzed 91,678 conversations and conducted six interviews with industry experts. Building on related research, we were able to identify and calculate criteria and metrics for evaluating customer-chatbot conversations. With our DSR project, we aim to support chatbot developers and operators leverage the potential of their data from customer-chatbot conversations (Følstad and Brandtzaeg, 2017). Ultimately, we aim to help ensure that experiences with customer service chatbots do not fall short of customer expectations (Bitner et al., 2010) by providing design knowledge for conversation mining systems that automatically analyze chatbot data at a conversational and process-oriented level to generate valuable insights for managers and developers.

Theoretical Foundations and Related Work

Chatbots in Customer Service

The customer service encounter that was once technology-free has become increasingly technologized through advances in information and communications technology (Fitzsimmons et al., 2018). Examples include electronic order entry, applications for price comparisons, or digital product configurators (De Keyser et al., 2019). Customers often no longer interact with human frontline service employees (FSE), but exclusively or first with technology. Either the technology can then finally process the customer request itself, or it forwards the interaction or results of it to a FSE (De Keyser et al., 2019; Kucherbaev, 2018).

A typical technology that augments and in some cases substitutes FSE are chatbots (Wirtz et al., 2018). Chatbots are suitable for customer service because they are always available, can save time, gather information, and add a more personal touch than web forms (Brandtzaeg and Følstad, 2017; Klopfenstein et al., 2017; Schuetzler et al., 2021). By responding questions or guiding users through process steps, chatbots in customer service typically assist customers at specific tasks (Schuetzler et al., 2021; Wirtz et al., 2018). For this reason, chatbots for customer service are technologically distinct from other conversational agents (McTear et al., 2016). For instance, intelligent assistants (e.g., Amazon's Alexa and Apple's Siri) and virtual companions (e.g., ELIZA) use artificial intelligence (AI) to understand and generate natural language, are often speech-based, have a broad range of topics, and a high number of interactions. Task-focused chatbots such as in customer service are mostly text-based, use AI only to classify user intention (intent), and have a limited range of topics and possible dialogues (Grudin and Jacques, 2019; Schuetzler et al., 2021). The data and metrics that describe the performance of these task-focused chatbots are therefore also different from those that characterize voice bots, where, for example, voice pitch and speaking speed play a stronger role (Aneja et al., 2020). Figure 1 shows the typical flow of a conversation with a text-based customer service chatbot. The communication takes place in an encoding-decoding sequence as known by the Shannon and Weaver model (Shannon and Weaver, 1949; Sperber and Wilson, 1996). As Figure 1 illustrates, the intention of a customer to contact the customer service with a concern **(1)** is translated in a concrete linguistic expression **(2)** when a chatbot is used. The chatbot then classifies this user message, e.g. "I want to cancel my order", to the intent underlying the statement **(3)**, using the NLU (Kucherbaev et al., 2018; McTear et al., 2016). According to the classified intent, the chatbot retrieves a predefined response **(4)**, as chatbots used in customer service use AI mostly for language understanding but rarely for generation (Schuetzler et al., 2021). In addition, after intent classification, chatbots can also launch dialogues **(5)** that are used to further specify the intention or to query or output data in connection with a backend (Kvale et al., 2020). Since these dialogues and responses are standardized, customer service chatbots are used especially for repetitive standard requests of customers (Thomaz et al., 2020).

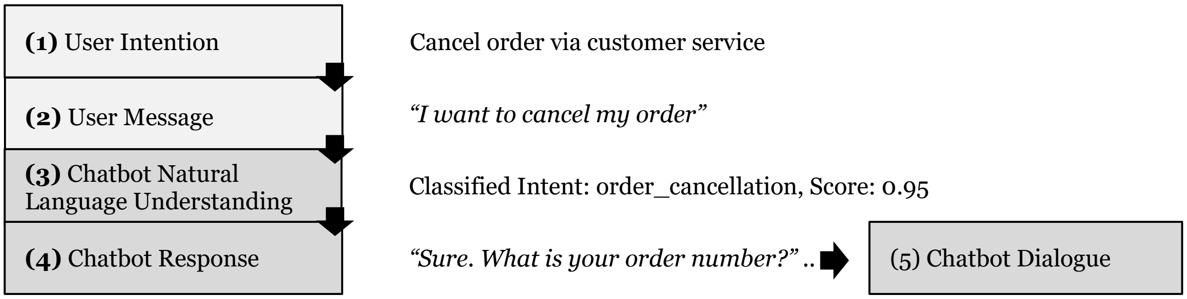
(1) User Intention
(2) User Message
(3) Chatbot Natural Language Understanding
(4) Chatbot Response
Cancel order via customer service
"I want to cancel my order"
Classified Intent: order_cancellation, Score: 0.95
"Sure. What is your order number?" ..
(5) Chatbot Dialogue

Figure 1. Flow of customer-chatbot conversations (based on Kucherbaev et al., 2018)

In order for chatbots to be profitable and realize their potential to automate standard requests in customer service, they must be used sufficiently by customers. However, failed service encounters lead to customer dissatisfaction and eventually complaints, poor word of mouth, inertia or switching behavior (Zeelenberg and Pieters, 2004). Thus, failed customer service encounters with chatbots lower future likelihood of use, especially as the choice of using chatbots competes with FSEs (Chen et al., 2022; Mozafari et al., 2022; Sands et al., 2021).

For this reason, there has recently been an increased focus on chatbot service quality and its measurement in service research. Existing research has predominately focused on extracting relatively simple metrics (e.g., technical metrics such as language model scores or user-related metrics such as conversation duration), manually coding or evaluating conversations (e.g., labeling the quality of chatbot answers), and performing interviews with chatbot users (e.g., to find out quality characteristics). For example, Chen et al., (2022) and Noor et al., (2022) found that the service quality of a chatbot increases when it is more human-like, adaptive and personal, closely collaborating with the FSE and continuously improved. For customer service chatbots, however, efficiency is particularly important (Chen et al., 2022; Prentice and Nguyen, 2021). Empirical studies have shown that customers place particularly high value on functional criteria for task-oriented chatbots, such as the error-free and efficient handling of their service requests (Følstad and Skjuve, 2019; L. Li et al., 2021; Meyer-Waarden et al., 2020). To guide customer service managers and chatbot developers in the deployment and design of chatbots when addressing these challenges, usage data is considered a valuable basis for decision-making (Bitner et al., 2010; Følstad and Taylor, 2021). It can provide customer service managers with insights into opportunities and challenges related to content and dialog design so that they can make informed decisions on how to bridge the gap between the chatbot's capabilities and customers' expectations and behaviors (Bitner et al., 2010; Chen et al., 2022). With automated data analysis, they would no longer have to base their decisions on high-level recommendations or time-consuming interviews. Instead, they can use a conversation mining system to generate decision-making guidance from usage data.

Conversation and Chatbot Analytics

Some part of chatbot research is specifically dedicated to the analysis of conversations, conversational breakdowns and repair strategies. A conversational breakdown occurs when a chatbot is unable to keep the conversation running smoothly, so in the worst case the customer service encounter fails and the customers leaves the conversation frustrated (Benner et al., 2021; Kvale et al., 2020). The first reason why such a breakdown can occur is an error in understanding the customer's intention (3). Customer's intention is not correctly understood when a non-classification (too low intent score) or a misclassification (wrong intent classified) occurs (Benner et al., 2021; C. H. Li et al., 2020). From the perspective of conversation theory, this can happen when an unintelligible message from the user (2) violates the cooperative principle of communication, e.g., by being verbose or unclear (Grice, 1975). Furthermore, a chatbot has problems of understanding if there is no "common ground" as a basis for the conversation between the customer and the chatbot (Clark, 1996). This happens when the customer is not informed on which topics the chatbot is trained or the chatbot covers a too small a range of topics (Kvale et al., 2020). The improvement opportunities for chatbot operators and developers regarding this problems are the expansion of topics, communication of chatbot capabilities and training of the language model (Beaver and Mueen, 2020). However, not all of these conversational problems can necessarily be filtered directly from the data. Low intent scores are immediately apparent as a breakdown, but misclassifications or insufficient responses (4) (for correctly hit intents) require more effort to determine (Kvale et al., 2020; C. H. Li et al., 2020). Ultimately, the breakdown may become apparent as the conversation progresses, for example, as the sentiment deteriorates, poor feedback or repetition of statements occur (Beaver and Mueen, 2020; Benner et al., 2021). Another reason for the failure of customer service encounters with chatbots is poorly designed dialogues (5) of service processes. Much of the chatbot analytics research has focused on conversational breakdowns regarding the language understanding of single user messages (Benner et al., 2021; T. J. J. Li et al., 2020). However, customer service chatbots often use fixed dialogues ("Please give me your contract number now") to interact with users (C. H. Li et al., 2020; Kvale et al., 2020). With our conversation mining system, we want to empower chatbot developers and operators to review customer service chatbot conversations completely, instead of only monitoring intent classification or aggregated metrics. Since dialogues are processes, problems with them are deviations from an optimal path. For example, if a customer service process in the dialogue consists of 3 consecutive user-side inputs, a deviation in the form actions, loops, or inputs of incorrect formats can indicate problems with the dialogue (Takeuchi et al., 2007; Yaeli and Zeltyn, 2021). Technically, for such process analysis, the raw data (row-by-row conversation events with timestamps) must be enriched with additional attributes (e.g., a counter indicating how often a user statement was repeated (Beaver and Mueen, 2020; Følstad and Taylor, 2021)). Then, for a selection of conversations, progressions and error types can be compared and grouped together (Quafari and Van der Aalst, 2012). In this way, we plan to make it possible for chatbot developers and operators to compare their expected conversations progresses and processes with the actual course and, based on this comparison, to adapt the design of the dialogues in a user-oriented manner (Van Der Aalst, 2012). Service quality can be increased and failed service encounters reduced, therefore increasing the likelihood of using chatbots again (Meyer-Waarden et al., 2020).

Method

Design Science Research

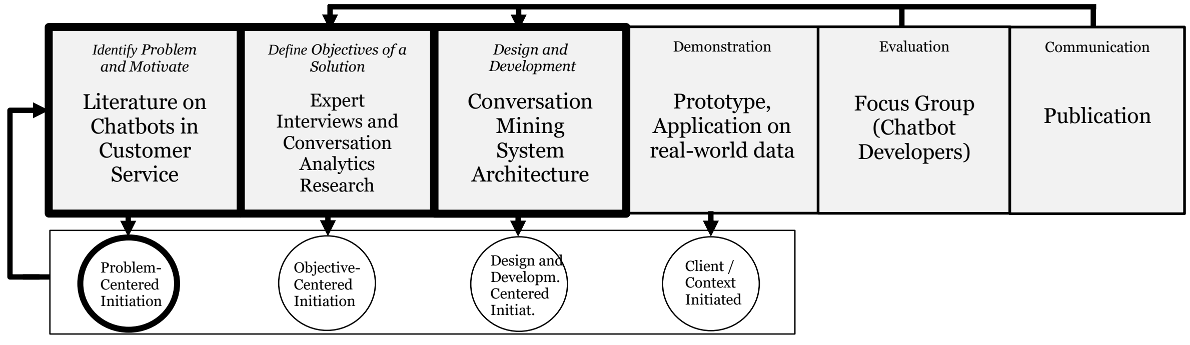
Identify Problem and Motivate
Literature on Chatbots in Customer Service
Define Objectives of a Solution
Expert Interviews and Conversation Analytics Research
Design and Development
Conversation Mining System Architecture
Demonstration
Prototype, Application on real-world data
Evaluation
Focus Group (Chatbot Developers)
Communication
Publication
Problem-Centered Initiation
Objective-Centered Initiation
Design and Developm. Centered Initiat.
Client / Context Initiated

Figure 2. Research Design – Cycle 1 (based on Peffers et al., 2007)

To assist chatbot operators and developers in assessing and improving the performance of customer service chatbots, our DSR project addresses the research question of how to design a conversation mining system. We consider the DSR approach to be particularly suited, since it involves iterative development and evaluation phases to ensure both relevance and rigor (Gregor and Hevner, 2013). Our DSR project follows the framework of Peffers et al. (2007), consisting of two design cycles, the first illustrated in Figure 2. In Cycle 1, we design a first prototype of the conversation mining system based on related research, expert knowledge, and log data which finally can be applied to real world conversations. The DSR project is conducted together with an industry partner that develops chatbots for energy industry customers (B2B) and provides us with conversational data of different chatbots as well as access to the chatbot operators (B2C) (Cardoso et al., 2015). The starting point of our DSR project was the initial problem that chatbot developers and chatbot operators, especially customer service managers, demand insights into the performance of their chatbots. To first better understand and frame the problem, the project started with a literature recap on the application of chatbots in customer service. In a second step, expert interviews were conducted with 4 chatbot developers of our industry partner and 2 experts from a chatbot operator. The first had the roles of product owner (3 years work experience), head of operations (3 years), user experience designer (1 year) and natural language understanding expert (0.5 years). We asked them in semi-structured individual interviews of 1.5 hours what they consider a successful and non-successful customer service chatbot conversation, why conversations fail, and what characteristics performance in conversations has. These interviews were recorded, transcribed, and analyzed via open coding. In addition, we interviewed 2 digitization strategy managers of a B2C energy company. They have 0.5 and 2 years of experience as chatbot operators. We discussed the current reporting and their analytical requirements in a 1.5 hours focus group, resulting in a requirement sheet. In addition, we had access to data from 91,678 real world customer service conversations of 40 different chatbots. Based on the structure and quality of the data and the methods of conversation analysis, a first design of a conversation mining system was proposed, which we present in this paper. As our next steps, we plan to finalize the prototype that instantiates our design. It will be evaluated with a larger focus group of chatbot developers (N = 10). In Design Cycle 2, we plan to refine the conversation mining system and conduct a second evaluation with a larger number of chatbot operators. By leveraging interaction information hidden in the data using user requirements from the customer service context and insights from chatbot analytics research, we aim to create design knowledge for conversation analytics tools that reveal chatbot performance in customer service (Gregor and Hevner, 2013).

Data

N messages/conversations	291,162/91,678
Basic data	User/Bot Messages and Forms Sent, Top Intents, Top Intent Scores, Timestamps
Additional Data	Frontend Interactions, Feedbacks (Message level), Clicked Links

Table 1. Metrics of provided data set

The first dataset which informed the development of the analytics backend, includes 91,678 conversations. In these, 291,162 messages were recorded, which with 3.18 messages per conversation indicates rather short task-related conversations. 6,411 of the 91,678 conversations are from the chatbot operator interviewed in the objective definition phase. As Table 1 shows, typical data of intent-based chatbots are available to us (C. H. Li et al., 2020), on the one hand the messages themselves, furthermore data of the intent classification as well as timestamps at all events. In addition, interactions with the frontend, feedback at the message level, and clicked links were also logged. No data for single customer recognition is available.

Designing a Conversation Mining System

Problem Identification and Motivation

Chatbots are particularly suitable for deployment in customer service and for handling standardized processes (Brandtzaeg and Følstad, 2017, 2018). However, they are only profitable for operators if they are used by customers in a larger scale. Therefore, customer service managers can limit the choice of channels for customers to address their requests. In this case, the service quality of the chatbot must be high enough to avoid failed customer service encounters that negatively impact the overall perception of a product or the company (Chen et al., 2022; Zeelenberg and Pieters, 2004). Or customers have a choice, in this case, the service quality of the chatbot must be so high that customers deliberately decide to use a chatbot based on recommendations and positive experiences, besides individual factors such as attitudes (Meyer-Waarden et al., 2020). Since customer service encounters tend to be very short and task-focused, evident in the few messages exchanged in customer chatbot conversations, performance aspects play a big role (Brandtzaeg and Følstad, 2017; Meyer-Waarden et al., 2020). Customers expect fast, simple and smooth processing when they cancel orders, change master data or have inquiries about products. If this is guaranteed, for routine tasks a chatbot is preferred over contact with a FSE. However, as research and real-world usage data have shown, customer-chatbot conversations repeatedly run into problems or break down (Aneja et al., 2020). This is due to 1. problems in language understanding or 2. problems with the chatbot's responses or dialogues (Benner et al., 2021). The patterns and indicators of these failed conversations vary and, while easily seen by human judgment, are difficult to capture automatically (Beaver and Mueen, 2020; Kvale et al., 2020). Because of this and the novelty of the technology, there is a lack of knowledge for automated assessment of chatbot performance in whole conversations, mass detection of problematic conversations, and operationalization of performance indicators (Beaver and Mueen, 2020; Følstad and Taylor, 2021).

Objectives of the System

The conversation mining system should provide chatbot operators and developers with data-driven insights about conversations in order to help them continuously assess and improve the performance of customer service chatbots. It requires two major components: 1) A frontend that provides access to the data in the form of visualizations that allows exploratory analysis of conversations. 2) A backend that preprocesses and stores the logged conversation data according to the defined requirements. To further understand the user requirements, we conducted interviews with chatbot operators and developers. Our interviewees were four experts from the industry partner, which is developing chatbots for the energy industry. They commented that, according to the depicted flow in Figure 1, a successful conversation includes understanding the concern, an appropriate response, and a conclusion to the process that was started. A failed conversation, on the other hand, is characterized by problems in natural language processing as well as breakdowns, technical problems, a stagnant dialogue flow, and inappropriate responses. Reasons for such problems that were mentioned were a possible mismatch between expectations and performance, e.g. in the form of missing but expected memory functions or backend connections of the chatbot, technical complexity and mentioning of topics that are outside the content spectrum, user messages that are too long or ambiguous on a linguistic level, as well as poor NLU training and non-user-centric design of dialogues (Følstad and Taylor, 2021). These causes should be reflected in a conversation mining system. The chatbot operators mentioned in our interviews that they would like to have visibility into which concerns the bot was able to process completely and which were not. They would like to see a topic-wise distribution of what the bot can handle well or not yet and a better insight into negatively rated conversations. In particular, they were also interested in gaining insight into which conversations had a handover to a human employee that can be triggered by typing "I want to talk to a human" and clicking a link. In this way, they want an accurate view on performance to reflect on the usefulness and profitability of the chatbot and to have an informed opinion on the maintenance of chatbot content and thematic language model training. The conversation mining system therefore will consist of filters, key metrics, actual conversation views, and visuals.

Design and Development

Based on the goals and assessments of the experts, in the third step of our DSR project, we developed an initial framework (Table 2) of metrics related to the performance of a customer service chatbot. The metrics are primarily derived from the literature on chatbot analytics (see column "Source"). We have split these metrics for success/failure of a conversation analogously to the established stages of communication to 1. user message, 2. natural language understanding, 3. chatbot response, 4. dialogue progress and 5. user (dis)satisfaction (Shannon and Weaver, 1949; Sperber and Wilson, 1996). In addition, referring to the typical data mentioned in chatbot analytics research (C. H. Li et al., 2020; Yaeli and Zeltyn, 2021) and the chatbot usage data we were provided with, we have explored and noted what data needs to be used in order to develop these metrics (see column “Data”). As we discovered, in terms of automatability, some metrics can be calculated easily or with moderate effort from the raw data, some must be calculated using external tools or APIs during data logging or ex post, and others require elaborate manual coding or subjective evaluation (see column "Automatability"). In the process of building our analytics pipeline, we excluded the metrics that require manual coding and focused on the metrics that can be generated automatically. Simple metrics (e.g., number of conversations) are also part of the conversation mining system, but do not require deeper exploration. As a conceptual basis, our framework guides the data pre-processing in our backend as well as the design of the frontend of the conversation mining system.

The first two conversational stages (user message and natural language understanding metrics) can be evaluated at the level of individual user messages (and the corresponding intent and response). The chatbot, response, the progress in the dialogue or dissatisfaction, on the other hand, refer to the context of the conversation. All metrics serve either as filter or as aggregated measures when reviewing the conversations on the frontend. Starting with user messages, a chatbot conversation can fail early on because the messages are too long, too short, or linguistically incorrect (Kvale et al., 2020, Reinkemeier and Gnewuch, 2022). The former can be checked quickly, for linguistic correctness an external spell checker is needed. Regarding the NLU, conversations with low or similarly high intent classification scores are easily filterable, mismatches as well as inputs beyond the known intents ('out-of-scope') can only be detected manually since the chatbot database does not know what it does not know (Beaver and Mueen, 2020; C. H. Li et al., 2020). The same applies to the relevance and understandability of the chatbot response, which also would have to be assessed manually (Følstad and Taylor, 2021), and are therefore not part of the automated system. If a user message is recognized correctly and a dialogue starts, there are various metrics describing dialogue progress. The repetition of user messages, intent hits, responses or certain events (for example, if a form appears again and again) as well as a clarifying statement by a user are signs of problems in the dialogue. These metrics do not require manual rework, but they do require an aggregation on the conversation level (Beaver and Mueen, 2020). Lastly, aborts of dialogues, entire conversations as well as handovers to the employee show that there was a problem. With the help of these conversation level metrics, the deviation of the smooth and actual course can be determined (Quafari and Van der Aalst, 2012). Last, user (dis)satisfaction in the form of direct feedback, sentiment, or escalation requests ("You're stupid, I want to talk to a human") attests the failure of the conversation (Akhtar et al., 2019). Based on this framework, the backend of our conversation mining system processes each conversation calculating all metrics that can be determined automatically.

	Metrics	Data	Automatability	Source
1. User message	Longevity Brevity Linguistic Correctness	User Message (Length) User Message (Length) User Message (Spell Check)	High High API	(Beaver and Mueen, 2020) (Reinkemeier and Gnewuch, 2022) (Kvale et al., 2020)
2. Natural Language Understanding	Low Score Ambiguity Mismatch Out-of-Scope	Top Intent Classification Score N Intent Classification Scores - -	High High Low/Indirect Low/Indirect	(Beaver and Mueen, 2020) (Benner et al., 2021) (C. H. Li et al., 2020) (Sperber and Wilson, 1996) (T. J. J. Li et al., 2020)
3. Chatbot Response	Relevance Response Understandability Response	- -	Low/Indirect Low/Indirect	(Aneja et al., 2020) (Følstad and Taylor, 2021)
4. Dialogue Progress/ Chatbot/User Repair	Intent Repetition Chatbot Response Repetition User Clarification User Message Repetition User Event Repetition Dialogue Abort Conversation Abort Handover → Deviation optimal dialogue	Top Intents Chatbot Responses User Message (Classification) User Messages (Similiarity) User Events Last Logged Event / Dialogue Last Logged Event / Conversation Handover Event Combination	Medium Medium Medium Medium Medium Medium Medium Medium	(Aneja et al., 2020) (Beaver and Mueen, 2020) (Følstad and Taylor, 2021) (Kvale et al., 2020) (C. H. Li et al., 2020) (Yaeli and Zeltyn, 2021) (Quafari and Van der Aalst, 2012)
5. User (Dis)satisfaction	Feedback Sentiment Escalation Request	Feedbacks (Message Level) User Message (Sentiment) User Message (Classification)	High API Medium	(Akhtar et al., 2019) (Beaver and Mueen, 2020)

Table 2. Chatbot performance metrics, data and automation levels

Conclusion and Next Steps

This paper provides first insights into the idea and design of our proposed conversation mining system. Based on in-depth exploration of chatbot conversation data from customer service and expert interviews, we identified the relevant (log data) metrics for successful and unsuccessful customer-chatbot interactions and designed an initial backend (analytics pipeline) that computes them. We also aim to design a user-friendly frontend for the conversation mining system as well and subsequently evaluate it with several chatbot operators to assesses the usefulness of the selected metrics and the visual design. We hope to increase their ability to gain insights into chatbot performance and take appropriate actions to increase the service quality of chatbot interactions (Chen et al., 2022), which we also plan to analyze in the context of a case study with one chatbot provider. Additionally, we plan to broaden our literature review on chatbot evaluation metrics to also include approaches that are not data-driven. Finally, our project is not without limitations. We specifically focus on intent-based task-focused chatbots in only one industry. Future research could extend the scope of conversation mining systems to other industries, other types of chatbots, and other types of chatbot data. Furthermore, we do not differentiate between individual users in the data, since no unique features are recorded except for the conversation number. Nonetheless, future research could expand upon our design to integrate additional data into the conversation mining system.

References

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

Aneja, D., McDuff, D., and Czerwinski, M. 2020. "Conversational Error Analysis in Human-Agent Interaction". *Proceedings of the 20th ACM Conference on Intelligent Virtual Agents*, pp. 1-8.

Beaver, I., and Mueen, A. 2020. "Automated conversation review to surface virtual assistant misunderstandings: Reducing cost and increasing privacy," *34th AAAI Conference on Artificial Intelligence*, pp. 13140–13147.

Benner, D., Elshan, E., Schöbel, S., and Janson, A. 2021. "What do you mean? A review on recovery strategies to overcome conversational breakdowns of conversational agents," *ICIS 2021 Proceedings*.

Bitner, M. J., Zeithaml, V. A., and Gremler, D. D. 2010. "Technology's Impact on the Gaps Model of Service Quality," in: *Handbook of Service Science*, P. P. Maglio, C. A. Kieliszewski, and J. C. Spohrer (Eds.), Cham: Springer, pp. 197–218.

Brandtzaeg, P. B., and Følstad, A. 2017. "Why people use chatbots," *Lecture Notes in Computer Science*, *10673 LNCS*, pp. 377–392.

Brandtzaeg, P. B., and Følstad, A. 2018. "Chatbots: Changing User Needs and Motivations," *Interactions*.

Cardoso, J., Fromm, H., Nickel, S., Satzger, G., Studer, R. and Weinhardt, C. 2015. "Fundamentals of Service Systems". In *Service Science: Research and Innovations in the Service Economy,* Issue 1.

Chen, Q., Gong, Y., Lu, Y., and Tang, J. 2022. "Classifying and measuring the service quality of AI chatbot in frontline service," *Journal of Business Research*, *145*, pp. 552–568.

Clark, H. H. (1996). *Using Language*. Cambridge University Press.

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

De Keyser, A., Köcher, S., Alkire (née Nasr), L., Verbeeck, C., and Kandampully, J. (2019). "Frontline Service Technology infusion: conceptual archetypes and future research directions," *Journal of Service Management*, *30*(1), pp. 156–183.

Fitzsimmons, J. A., Fitzsimmons, M. J., and Bordoloi, S. 2018. *Service Management: Operations, Strategy, Information Technology*, McGraw-Hill Education (ed.), 9th Edition.

Følstad, A., and Brandtzaeg, P. B. 2017. "Chatbots and the new world of HCI," *Interactions* (*24*:4), pp. 38–42.

Følstad, A., and Skjuve, M. 2019. "*Chatbots for customer servic*e: user experience and motivation," *CUI '19: Proceedings of the 1st International Conference on Conversational User Interfaces,* pp. 1–9.

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

Gao, M., Liu, X., Xu, A., and Akkiraju, R. 2021. "Chatbot or Chat-Blocker: Predicting Chatbot Popularity before Deployment," *DIS 2021 - Proceedings of the 2021 ACM Designing Interactive Systems Conference: Nowhere and Everywhere*, pp. 1458–1469.

Gregor, S., and Hevner, A. 2013. "Positioning and Presenting Design Science Research for Maximum Impact," *MIS Quartely* (37:2), pp. 337–356.

Grice, P. 1975. Logic and Conversation. In *Speech Acts*. Brill, pp. 41–58.

Grudin, J., and Jacques, R. 2019. "Chatbots, humbots, and the quest for artificial general intelligence," *Conference on Human Factors in Computing Systems - Proceedings*, pp. 1–11.

Hill, J., Randolph Ford, W., and Farreras, I. G. 2015. "Real conversations with artificial intelligence: A comparison between human-human online conversations and human-chatbot conversations," *Computers in Human Behavior*, *49*, pp. 245–250.

Klopfenstein, L. C., Delpriori, S., Malatini, S., and Bogliolo, A. 2017. "*The Rise of Bots,*" Proceedings of the 2017 Conference on Designing Interactive Systems, pp. 555–565.

Kucherbaev, P., Bozzon, A., and Houben, G. J. 2018. "Human-Aided Bots," *IEEE Internet Computing* (22:6), pp. 36-43.

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

Li, C. H., Yeh, S. F., Chang, T. J., Tsai, M. H., Chen, K., and Chang, Y. J. 2020. "A Conversation Analysis of

Non-Progress and Coping Strategies with a Banking Task-Oriented Chatbot," *Conference on Human Factors in Computing Systems - Proceedings*, pp. 1–12.

Li, L., Lee, K. Y., Emokpae, E., and Yang, S. B. 2021. "What makes you continuously use chatbot services? Evidence from chinese online travel agencies," *Electronic Markets* (*31*:3), pp. 575–599.

Li, T. J. J., Chen, J., Xia, H., Mitchell, T. M., and Myers, B. A. 2020. "Multi-modal repairs of conversational breakdowns in task-oriented dialogs," *UIST 2020 - Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pp. 1094–1107.

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

Meyer-Waarden, L., Pavone, G., Poocharoentou, T., Prayatsup, P., Ratinaud, M., Tison, A., and Torné, S. 2020. "How Service Quality Influences Customer Acceptance and Usage of Chatbots?," *Journal of Service Management Research*, *4*(1), pp. 35–51.

Mozafari, N., Weiger, W. H., and Hammerschmidt, M. 2022. "Trust me, I'm a bot – repercussions of chatbot disclosure in different service frontline settings," *Journal of Service Management*, *33*(2), pp. 221–245.

Noor, N., Hill, R., and Troshani, I. 2022. "Developing a Service Quality Scale for Artificial Intelligence Service Agents," *European Journal of Marketing*, ahead of print.

Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. 2007. "A design science research methodology for information systems research," *Journal of Management Information Systems*, *24*(3), pp. 45–77.

Prentice, C., and Nguyen, M. 2021. "Robotic service quality – Scale development and validation," *Journal of Retailing and Consumer Services*, *62* (June), pp. 1-7.

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

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.

Reinkemeier, F. and Gnewuch, U. 2022 "Designing Effective Conversational Repair Strategies for Chatbots," *ECIS 2022 Research Papers*.

Sands, S., Ferraro, C., Campbell, C., and Tsao, H. Y. 2021. "Managing the human–chatbot divide: how service scripts influence service experience," *Journal of Service Management* (*32*:2), pp. 246–264.

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

Shannon, C. E., and Weaver, W. 1949. The Theory of Mathematical Communication. *International Business*, 131.

Sperber, D., and Wilson, D. 1996. *Relevance: Communication and Cognition* (2nd). Wiley-Blackwell.

Takayama, J., Nomoto, E., and Arase, Y. 2019. "Dialogue breakdown detection robust to variations in annotators and dialogue systems," *Computer Speech and Language*, *54*, pp. 31–43.

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.

Thomaz, F., Salge, C., Karahanna, E., and Hulland, J. 2020. "Learning from the Dark Web: leveraging conversational agents in the era of hyper-privacy to enhance marketing," *Journal of the Academy of Marketing Science* (*48*:1), pp. 43–63.

Van Der Aalst, W. 2012. "Process Mining : Overview and Opportunities," *ACM Transactions on Management Information Systems* (*3*:2), pp. 1–17.

Van Doorn, J., Mende, M., Noble, S. M., Hulland, J., Ostrom, A. L., Grewal, D., and Petersen, J. A. 2017. "Domo Arigato Mr. Roboto: Emergence of Automated Social Presence in Organizational Frontlines and Customers' Service Experiences," *Journal of Service Research* (*20*:1), pp. 43–58.

Wirtz, J., Patterson, P. G., Kunz, W. H., Gruber, T., and Paluch, S. 2018. "Brave new world : service robots in the frontline world," *Journal of Service Management* (*29*:5), pp. 907–931.

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.

Zeelenberg, M., and Pieters, R. (2004). "Beyond valence in customer dissatisfaction: A review and new findings on behavioral responses to regret and disappointment in failed services," Journal of Business Research (57:4), pp. 445–455.