2019 IEEE 21st Conference on Business Informatics (CBI)

The Potential of Chatbots: Analysis of Chatbot Conversations

Mubashra Akhtar
TU Wien
Research Unit of E-Commerce
Vienna, Austria
mubashra.akhtar@ec.tuwien.ac.at

Julia Neidhardt
TU Wien
Research Unit of E-Commerce
Vienna, Austria
julia.neidhardt@ec.tuwien.ac.at

Hannes Werthner
TU Wien
Research Unit of E-Commerce
Vienna, Austria
hannes.werthner@ec.tuwien.ac.at

***Abstract*—**
The idea of utilizing computers for question answering tasks has been around from the early beginning of these systems. First algorithms with the aim to accomplish this were already implemented in the early 1960s. In recent years, chatbots have been gaining enormous popularity in various fields. In the context of business applications, they are considered as useful tools for improving customer relationships. In this paper, chat conversations between customers and the chatbot of a telecommunication company are analysed to find out if these interactions can be used to determine a) users' topics of interests and b) user satisfaction. To reach this goal, chat conversations are interpreted as sequences of events and user inputs are analysed with the help of text mining techniques. The study shows that based on users' written conversational contributions, valuable insights on users' interests and satisfaction can be gained. The majority of users leave the chat conversation after a short period of time if the chatbot was not able to give the desired answer right away. Moreover, a huge number of conversations deal with similar topics. Our results imply that companies offering chatbots must thoroughly analyse the collected data to gain more insights into their customers' needs. Based on our findings, they can improve customers' satisfaction by offering personalized service and implementing real-time feedback.

***Index Terms*—Chatbots, Text Mining, Event Sequence Analysis, Network Analysis, Telecommunication**

I. Introduction

Many companies have recognized chatbots as "the next big thing" in terms of customer relationships. In today's digital age these relationships are shaped by an empowerment of customers due to increased information availability, digital communication channels and more diverse possibilities for reaching customers [1]. Regarding the usefulness of chatbots from the customers' point of view, a chatbot enables 24-hour customer service, personalized interaction and no waiting time. For companies, chatbots entail time and cost savings as many processes can be automated and employees can be appointed to more complex tasks [1].

Huge and famous information technology companies such as Google, Facebook and IBM have contributed to chatbot development and research in recent years. One of the most well known examples is the question answering system named Watson by IBM. Further, mentionable examples are "Siri", which is developed by Apple, and Amazon's "Alexa". Both systems can understand text as well as speech [2].

However, almost all of these developments and research activities concentrate on chatbots using the English language. In a recent study [1], German stock companies have been analysed in regard of chatbots. Out of 80 companies, 12 companies (15%) use a chatbot for customer communication. This implies, that the market for non-English speaking chatbots also exists and is on the rise.

The goal of our study is to examine interactions between users and chatbots. We want to find out, in particular, whether the analysis of these interactions can lead to additional insights regarding the users' information needs. First of all, the aim is to investigate whether users are able to find the answers they are looking for while contacting a chatbot. Secondly, we want to examine whether they are satisfied with the received answer. This study also indicates that chat conversations between customers and chatbots are valuable sources of information and useful for improving customer satisfaction. The results are applicable for all companies, which have implemented chatbots, regardless of their business area.

Consequently, our chatbot analysis concentrates on the following research questions:

- RQ1: Are chat conversations between users and chatbots sufficient sources of information to determine users' topics of interests?
- RQ2: How can chat conversation be used to determine if users are satisfied after the chat and their needs are fulfilled?

To answer these questions, text mining and event sequence analysis have been chosen as appropriate techniques. More precisely, users' topics of interests are determined by frequent term extraction and analysing bigrams and trigrams, which occur in conversations. Users' satisfaction after chatting is determined by analysing their feedback comments using text mining. Furthermore, conversations are transformed to sequences of events, which are subsequently modelled with the help of network analysis, and the quality of the answers is predicated from the moment users leave the conversation.

978-1-7281-0650-2/19/$31.00 ©2019 IEEE
DOI 10.1109/CBI.2019.00052

The rest of the paper is organized as follows: In the next section we introduce some theoretical concepts related to our study. Section III gives an overview of the available data and introduces the methods that are used. In section IV the results and findings of the study are presented, and they are discussed in section V. Furthermore, their applicability for businesses offering a chatbot for their users is discussed. Finally, in the last section we conclude the entire study, describe limitations and outline future work.

II. BACKGROUND

The use of computers for answering textual questions goes back to the early 1960s, when systems implementing question answering algorithms were first built [3]. In this section common paradigms used by computational systems for question answering as well as different types of dialog systems are presented. Finally, an introduction to industrial chatbots is given and the telecommunication chatbot, which is analysed in this work, is introduced.

A. Question Answering Paradigms

1) Information-retrieval-based Question Answering: Information-retrieval-based question answering systems aim to respond to users questions by finding short texts in a collection of documents available to the system e.g. the web, which contains possible answers to the proposed questions or a database. This paradigm strongly relies on information availability on the web (or in other systems) in form of a vast collection, which can be searched for the answer [3].
2) Knowledge-based Question Answering: Knowledge-based question answering systems (KB-QA) answer questions in natural language using a structured database. The database can be either a full relational database or a simpler database. For frequently asked questions, rule-based methods are convenient as simple rules can be written for questions that occur often. Systems using a rule-based method typically utilize a knowledge base consisting of facts and rules.

B. Types of Dialog Systems

Dialog systems, which are also known as conversational agents, are systems designed for communicating with users using natural language. These systems can be sub-categorized into two classes: task-oriented dialog systems and non-task-oriented dialog systems [3].

1) Task-oriented dialog agents are systems built for a certain purpose and for conducting short conversations. Well-known digital agents e.g. "Siri" and "Alexa" are task-oriented dialog agents, which are designed for simple tasks such as making phone calls, describing routes or finding restaurants. Conversational agents, which are installed on companies' websites to assist customers with their problems and to answer their questions, are usually also task-oriented dialog agents [3]. These agents are often based on a certain domain ontology.
2) Non-task-oriented dialog agents are used for longer and more complex interactions with the purpose of imitating conversations between humans. These systems don't focus on a certain task but are meant for entertaining users. "Xioaice" developed by Microsoft Peking is an example of a non-task-oriented dialog agent [2]. The chatbot is designed more like a friend. On the one hand she can answer questions e.g. about the weather or news and on the other hand she has the ability to react to the emotional states of individuals talking to her. Therefore, many users reported that they did not recognize in the first ten minutes of the conversation that it was a chatbot they are interacting with [4].

C. Application Context

The presented work focuses on the telecommunication domain. The data was provided by a large Austrian telecommunication company. This company has implemented a chatbot on its website, which welcomes users as soon as they visit the site and answers their questions. This chatbot can be classified as a task-oriented dialog agent, which aims to assist users during their search process on the company's website. Therefore the chatbot answers simple and common user questions directly on the one hand and directs them to other question answering channels on the other hand. The chatbot is a knowledge-based question answering system. Answers to users' questions are retrieved using different query languages to map users' questions to the matching answers. The company's aim behind implementing a chatbot, is to increase user satisfaction and to reduce costs by automating certain aspects of customer service. This study is a first attempt to find out how well it works.

III. METHODOLOGY

A. Analytics Process

The process applied for this work is based on the Cross Industry Standard Process for Data Mining, also known as the CRISP-DM reference model. The process models the entire life cycle of data mining projects and consists of six different phases: business understanding, data understanding, data preparation, modelling, evaluation and deployment. The phases are not executed sequentially but allow moving back and forth between them. [11].

B. Data Overview

The chatbot data used for analysis covers conversations between users and a telecommunication chatbot between February and August 2016. For each single month between February and August 2016, a separate datafile is provided, which captures the interactions of each particular month. The entire data from February to August 2016 is merged into one single dataset. The resulting dataset contains 673,176 entries. Each entry of the dataset represents an interaction i.e. a question asked by the user, an answer given by the chatbot or feedback provided by users. The data can be sub-categorized into sessions and interactions. Each interaction is allocated to

a session. In sum our dataset X captures 215,859 different sessions.

It is important to mention that due to the condition of data, analyses could only be done in a restricted way. A huge part of the available URLs were outdated and not available on the web anymore. Moreover, the user id, which is essential for analyzing user behavior over time, was not consistent. Therefore it was only possible to analyse user behavior session-wise, i.e., we can not follow a user across different sessions.

The interactions between users and the chatbot are described by 32 different attributes. For our analysis only a subset of these attributes was relevant. These can be grouped into three categories: basic attributes, attributes used for feedback analysis and attributes used for event sequence analysis. Basic attributes give an overview of the available data e.g. an unique ID for each interaction, an ID for each session in order to group interactions into session and the timestamps. The second group includes all attributes related to users' feedback, for example the feedback comments and scores. The last group counts attributes, which are used for the event sequence analysis. Three attributes belong to this category: the previous page of an interaction, the pre-labelled interaction type and the clicked URL.

Before going into more detail, an exemplary chatbot conversation is presented in Figure 1 This conversation begins at the starting page of the companys website. Analog to all other conversations, this chat has a unique session ID and it consists of five different interactions. First, a question is proposed by the user, which translates to "Can you assist me with the contract extension?". Next, the chatbot gives the answer that individual offers for contract extension can be found in the users' area when logging in with the personal account, and it provides a link to the log-in area. This gives the user the possibility to exit the conversation and find the answer on the website of the company. The user receives also the possibility to further specify the question by clicking on either "extension for smartphone contracts" or "extension for Internet contracts". The third interaction is user feedback given to the chatbot. The user has the possibility to rate the answer positively by choosing an upward thumb or negatively with a downward thumb. After choosing the downward thumb as the fourth interaction, the user is asked for a feedback comment at the end of the conversation. The fifth and final interaction is a feedback comment by the user: "the chatbot provided the wrong answer". Note that the user has the possibility to rate each interaction individually with a "like" or "dislike".

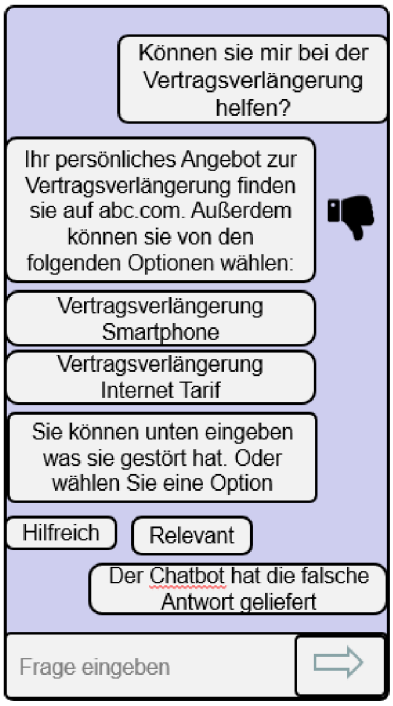

Fig. 1. An exemplary conversation between the chatbot and a user

C. Data Preprocessing & Analysis Approaches

TOPIC EXTRACTION

In order to apply text mining on chat conversations, the textual data has to be preprocessed first. The following preprocessing steps are performed:

- *Corpus generation*: Out of the conversations, text corpora are generated. A corpus is described in the linguistic sciences as a "body of written text or transcribed speech" [5]. In order to generate corpora, first of all those columns out of the dataset are selected, which include the textual data to be analysed. Two different corpora are created out of this data: a conversation corpus and a feedback corpus. The feedback corpus contains text related to feedback e.g. out of the feedback comments provided by the users. Next the steps are described, which are conducted on the corpora during data preprocessing.
- *Eliminating extra whitespace*: Extra whitespace between words is eliminated from the text in order to structure the text better.
- *Stopwords removal*: Next, a list of words, which are known as stopwords, is removed from the corpus. Stopwords are words that usually don't increase the predictive capability of a text, e.g. articles like 'the' or pronouns like 'us' [6]. The negating words 'nicht' and 'nein' are kept although they are categorized as stopwords because an elimination of these negotiations leads to distortion of the feedback analyses.
- *Lower case transformation*: The words are transformed to lower cases.
- *Tokenizing*: One of the initial steps of natural language processing is to identify tokens. During the process of tokenization, the entire text from the dataset is decomposed in tokens. These tokens can be e.g. single words or sentences [7]. The chatdata is tokenized in single words.

- *Stemming*: Stemming refers to the process of converting the identified tokens into a standard form. This step reduces the number of distinct types within a corpus as words such as 'types' and 'type' are counted as instances of the same type [6].
- *Creating term-document matrices*: Term-document matrices are matrices, which columns represent terms that occur within a corpus. Their rows correspond to a document of the corpus. The matrix displays the frequency of each single term that occurs within a specific document. In the term-document matrix generated out of the chat data, each interaction is one document of the corpus.

In contrast to classic data mining techniques, which are used to extract knowledge out of structured databases, text mining techniques have the aim to extract information out of unstructured textual data [8]. After preprocessing the conversations, sets of words, which occur together are analysed. This is done to determine topics occurring in these chats. A set of two successive words is also called 'bigram'. Analog a set of three words is named 'trigram'. The context of conversations might not be captured when the analysis is done at single word level. For example, the context of the phrase "mobile phone defect" is completely lost when each of the three words is viewed individually.

USER SATISFACTION

User satisfaction can be measured based on explicit and implicit feedback provided by a costumer; depending on whether users share their opinion on a certain product or service intentionally or unintentionally. To determine the satisfaction of the users after chatting, their explicitly provided feedback was analysed first.

In the dataset used for this study, explicit feedback is provided using feedback scores and comments. In sum the dataset includes 7,995 different feedback comments, i.e., although the users have the opportunity to provide feedback for most of the interactions, only very few of them were rated by the user. Concerning feedback scores, the data captures two different score ranges: First, in some chat sessions, users have the possibility to give feedback by choosing one score out of {-2, -1, 0, 1, 2}. The smallest score of '-2' is equal to one star on the chatbot's interface and connotes dissatisfaction. The highest score is '2' and it represents a high satisfaction with the chatbot's answers. Secondly, in some chat sessions a feedback scale, where users can only choose one out of two scores, is available: '-1' and '1' or 'like' and 'dislike' was used. The five-star scale is a Likert Scale, which was named after Rensis Likert [9] who introduced this technique for assessment tasks, is used to indicate how much the surveyed person agrees or disagrees, approves or disapproves to a given statement [10]. Scales consisting of two answer possibilities are known as dichotomous scales e.g. 'like' and 'dislike'. The different feedback scales are attributable to the fact that the dataset captures chatdata over seven months, from February to August 2016. During this period of time the cooperating telecommunication company changed the feedback options from a five-point scale to a two-point scale. The reason for that is not know to us.

Since only few interactions are explicitly rated, our goal is to assess user satisfaction also based on implicit feedback extracted from user behaviour in chat conversations. Therefore, we capture conversations as sequences of events, which are extracted out of the user sessions. Subsequently, these events are modelled as networks and analysed using network analysis techniques [16].

In Figure 2 an exemplary session and the related events, which are determined within that conversation, can be seen. The first event of this conversation is a question asked by the user: "Can you help me with the contract extension?". This question is labelled with "qa_734". This internal label is automatically assigned by the chatbot to categorize the input of the user into different predefined topics (e.g., invoice-related, mobile data, etc.) and to determine the reaction of the chatbot. A reaction of the chatbot is either a non-specific answer to inquire further the user's information need or a specific answer, where the user is pointed to a FAQ or where a link to another page is provided.

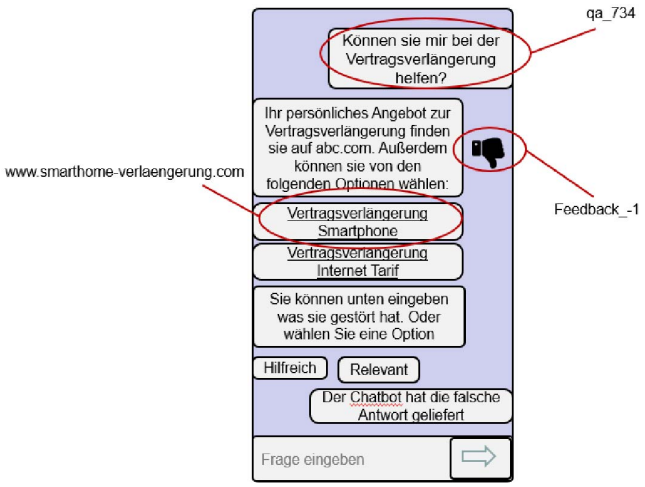

Fig. 2. Extracted events out of a conversation

Since each event, i.e., each interaction of the user as well as the reaction of the chatbot, receive an internal label, each session can be modelled as sequence of events represented by the corresponding labels. Therefore, when considering all the sessions together, a network of events can be constructed. Two nodes are connected, if the two events happened successively with the direction of the edge pointing from the first event to the second one. Weights assigned to the edges indicate how often two events take place consecutively in all interactions. Thus, this type of network allows a high-level view by summarising all sessions in the time period under consideration.

Converting all chat sessions of February to event sequences and presenting them within one network, we receive in total

3,181 nodes and 917,071 edges. In Figure 3 a subset of the extracted events of February are presented as a network. The question and non-specific answer nodes are coloured black in the networks. The red nodes are the specific answers that should provide the information the user is looking for, i.e., FAQ and link clicks. The orange nodes represent user feedback, i.e., when users providing feedback scores for a certain interaction. For the other months, this can be done accordingly.

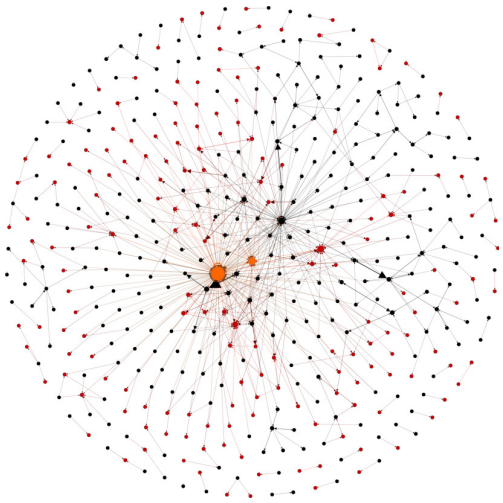

Fig. 3. Sub-network of events in February 2016

In the next step, a measure is defined to evaluate the quality of an answer provided by the chatbot.

It is desired by the telecommunication company that a user receives a good answer and leaves the chat afterwards (i.e., after the right FAQ or a good link is provided by the chatbot). In the system, there are specific labels for answers that are provided. Thus, nodes that represent such answers to user inquiries are supposed to have a low weighted out-degree $outd_w$ as compared to their weighted in-degree ind_w (i.e., fewer edges leaving that node compared to its incoming edges). On the other hand, nodes representing answers that are not of interest to the majority of users might have a quite high weighted out-degree as a lot of users continue their search and visit further nodes after reaching this node.

Based on this idea, we propose the following way to measure the quality q_n of an answer node:

$$q_n = ind_w/(ind_w + outd_w) = ind_w/d_w$$

Thus, nodes with q_n close to 1 represent answers after which many users stopped interacting with the chatbot, nodes with much lower q_n represent answers that typically lead to an on-going search process by the users. A score of 0.5 indicates that all users who reach the specific node, continued their search afterwards.

IV. Results

Before discussing the users' topics of interests and their satisfaction in more detail, the basic attributes are analysed in this section. These attributes provide an overview and basic understanding of the available data and allow grouping the data according to chat sessions, time, categories or users. The attributes interaction ID and session ID are essential in respect to grouping the data for the analysis.

The user ID is a hashed contract ID provided by a Single-Sign-On (SSO) token as soon as users sign in on the website. Identifying the users would allow us to track and analyse user behavior across different sessions but as this attribute is not consistently provided by our cooperation partner in the available dataset, identifying the users is not possible. Identifying users allows to consider background information of them e.g. age and contract details while discussing users' behavior, preferences and needs.

The timestamp attribute captures various aspects of sessions and interactions e.g. the duration of sessions, when the sessions start and end, as well as the distribution of sessions across weekdays. It facilitates analyzing the workload of the chatbot at different times of the day and different days of the week.

To analyse the amount of open sessions at a specific time of the day, the entire conversation data is aggregated according to the starting points of the conversations. In the entire dataset, a peak of chatbot interactions can be found between 9 am and 10 am. Moreover, the highest number of interactions does not take place in the evenings as one might suppose, but in the traditional office hours between 9 am and 5 pm. This can be related to the fact that during office time it is easier for a person to chat than to make a phone call and talk to a customer service employee. Furthermore, the distribution of chat sessions among the different days of the week were analysed. The biggest amount of interactions (18.2%) took place at the beginning of the week (Monday and Tuesday). The interactions on weekends are lower: 8.8% on Saturdays and 10.4% on Sundays.

Moreover, analysing the duration of conversations, it becomes apparent that the majority of users want an answer to the stated question quickly. Otherwise they leave the chat conversations very soon. In total 122,540 chat sessions out of 215,859 last for less than one minute as these conversations end within seconds.

TOPIC EXTRACTION

Table I displays a list of trigrams occurring in conversations. The topics *sim card locking* and *unlocking*, *the homenetbox*, *customer service* and *roaming* become apparent.

As the provided dataset is labelled with chat topics, the extracted topics using bi- and trigram generation can be compared to this data labelling and can be used to identify new topics not considered before. The company labelled a small part of all interactions with labels. Concretely 101,115

First Word	Second Word	Third Word	Frequency
sim	lock	aufheben	723
rechnung	einfach	erklärt	513
home	net	box	452
zonenroaming	für	vertragskunden	425
nummer	wurde	gesperrt	403
sim	karte	sperren	285
my	homenet	unlimited	276
neue	sim	karte	276
nein	kundenservice	kontaktieren	248
internet	funktioniert	nicht	238

TABLE I
FREQUENT TRIGRAMS IN THE CHAT DATA

out of 673,176 interactions have an assigned topic label and in total 64 different categories exist to which sessions can be matched. The categories with the highest amount of chat sessions are “Billing”, “Email” and “Homenetbox”. The category “Billing” has the biggest share of 6,575 chat sessions.

USER SATISFACTION

After preprocessing the textual data as described before, a word cloud of the most frequent terms within these comments was created. The word cloud can be seen in Figure 4. As it is mentioned before, feedback is often negatively related. This fact is commonly known as humans tend to give more often feedback when they are dissatisfied. The words ‘nicht’ and ‘nein’ are not removed during stopwords elimination because an elimination of these negotiations leads to a distortion of the feedback analysis. Related words to the term ‘not’ are ‘helpful’ with a correlation of 0.50, ‘suggested’ (0.48), ‘question’ (0.43), ‘became/was’ (0.39), ‘understood’ (0.37) and ‘correct’ (0.36). The word ‘answer’ is correlated to ‘defective’ (0.58), ‘question’ (0.27) and ‘wrong’ (0.15). The correlation is calculated based on the occurrences of the words within the same document.

Figure 5 summarizes and visualizes the feedback ratings. Out of 673,176 interactions, in 94.5% no feedback score is provided. The remaining 5.5% can be divided in 23% positive feedback and 77% negative feedback. This low number clearly proves that reliable ways to assess indirectly the satisfaction of the users have to be established.

When analysing feedback for the top three categories (“homenetbox”, “E-mail” and “billing”), it becomes apparent that feedback provided to chats dealing with the topic e-mail is more negative (87.8%) than for homenetbox (73.1%) and billing (79.4%). Such findings can help to improve the chatbot’s answers in clearly negatively rated categories.

Figure 6 displays a boxplot generated out of the answer quality values of the month of February. The minimum value of the boxplot is equal to 0.5. In this case the ingoing degree is equal to the outgoing degree and no chat conversation ends after this node. The maximum value of the boxplot is equal to one. Nodes having an answer quality of one, don’t have any outgoing edge as no user continued the chat conversation after reaching this node. As it can be seen in the graphic, the boxplot is right skewed and has a median (0.625) closer to the

Fig. 4. Frequently used terms in the corpus $c_{Feedback}$

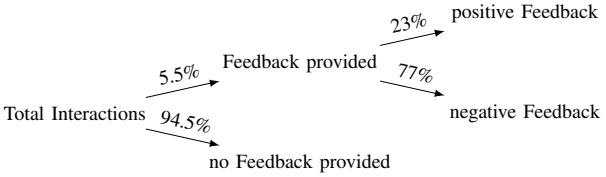

Fig. 5. Summary of feedback scores

first quantile (0.515) than to the third quantile (0.793). As it can be seen, half of all answers have a small quality between 0.5 and 0.625 compared to the entire scale (0.5-1).

A qualitative evaluation was conducted by going through the results in detail. In addition, both the topics obtained by text mining and the introduced definition of the quality of answer nodes were presented to the company in several meetings in order to find out if these approaches are useful to them and if reasonable conclusions can be drawn. In both cases this was confirmed by the company.

V. DISCUSSION & IMPLICATION OF RESULTS

In this work, we tried to answer the following questions:

- RQ1: Are chat conversations between users and chatbots sufficient sources of information to determine users’ topics of interests?
- RQ2: How can chat conversation be used to determine if users are satisfied after the chat and their needs are fulfilled?

The results obtained in the previous sections demonstrate that chat conversations are sufficient data sources for determining users’ interests. Furthermore, although user satisfaction cannot be assessed in a straightforward way, these interaction

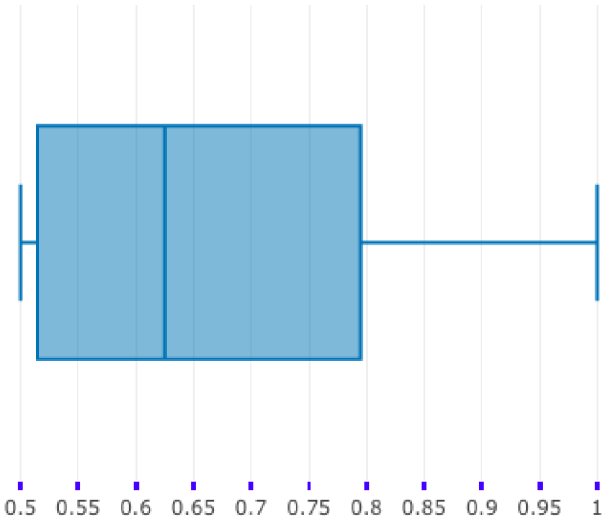

Fig. 6. Boxplot of answer qualities in February 2016

data provide ways to get deeper insights whether appropriate answers are delivered by the chatbot or not.

Every company offering a chatbot to its users should make use of this valuable source of information to improve its service, regardless of the business domain it operates within. Understanding the users, their needs, consumption patterns, behavior and preferences helps businesses to tailor their products, services and strategies according to their customers. Modelling, profiling, analyzing and understanding users becomes increasingly important in many different industries and a key to success in todays data driven world [12]. Based on the results of this study, the following recommendations are given for using chat conversation data to full extend:

Knowing users' topics of interests enables understanding their information needs and reacting to these needs fast. This knowledge of users' needs can be used to increase their satisfaction by helping them to find the right answer quickly and easily. Therefore we suggest to offer the most popular topics and questions right at the beginning of conversations for users to choose from. On the one side this will increase their satisfaction and on the other side, users don't have to write texts, which have to be understood and answered from the system. Regular mining of the users free text inputs can help to identify new and relevant topics not or not correctly considered by the chatbot.

As it has been demonstrated in this study, explicit user feedback is important for improving the chatbot's functionalities. The Likert scale, which the telecommunication company provided first, gives users the option of rating the chatbot on a scale from one to five stars. Later, the scale was adapted and changed to a dichotomous scale providing two options for feedback: an upward thumb (like) and a downward thumb (dislike). One star, two stars and a click on the dislike-button are interpreted as negative feedback. Whereas five and four stars, as well as clicking on the like-button is understood as positive feedback. Nevertheless it is important to mention, that human beings tend to give feedback more often when the feedback is negative than giving positive feedback. As mentioned above, this was also observed in our analysis. To support explicit feedback, short feedback scale should be chosen. In 37,065 interactions users provide feedback by choosing a feedback score. 8,162 interactions are 'liked' by users and 26,344 'disliked'. Considering the five-stars feedback scale, 2,201 interactions receive only one star, 67 interactions get two stars, 41 three stars, 61 four stars and 189 interactions get a score of five stars.

This forces users to take a position whether they are satisfied or not with an answer. Using a Likert scale with a middle option, allows users to choose a rating which expresses neither satisfaction, nor dissatisfaction.

The network of events as it is displayed in Figure 3 and the resulting quality measure for answer nodes, help extracting mistakenly installed question - answer pairs. A quality score of 0.5 indicates that all users who reach the specific node continue their search afterwards. This does not necessarily mean that the answer is completely wrong and the users are not satisfied. However, it indicates that something might be wrong since the users do not accept the provided answer in the way it was expected when designing the chatbot dialogues.

Thus, based on these calculated measures a targeted improvement of the chatbot is possible. Half of the answer nodes, which include links and answers given by the chatbot, have a small answer quality between 0.5 and 0.625. The small answer qualities give evidence that these answers are not matched appropriately to the questions of the users and the user is forced to continue the chat in order to get the right answer. The underlying assumption is that a user leaves the conversation when the right answer is found. Though, it is important to mention that a user might also stop the conversation due to frustration and not because the answer was provided.

The node with the highest answer quality of 1 is a link, which is related to the phonebook of the user. In twelve conversations users get this event as an answer to their question and in none of these twelve conversations the user continues to chat with the chatbot afterwards. The recommended website allows users to configure their telephone numbers in the phonebook. Having a look at the answer nodes with the smallest answer quality is even more interesting as this knowledge helps to identify and eliminate falsely installed answer nodes. Each of the 21 users who received a link related to direct debit as an answer, continue to chat. This fact implies that this answer node has to be reconsidered and if necessary, exchanged. Further nodes with a small answer quality, are website links which belong to the following topics: notice period, basic service packages, e-bill and additional service packages. In order to improve users' satisfaction, conversations which include nodes with small answer qualities have to be analysed in more detail. The falsely implemented answers have to be reconsidered and adjusted.

VI. Limitations & Future Work

In this study we showed various ways how interaction data with a chatbot can be used to enhance a companies' knowledge about the needs of their users as well as user satisfaction. This knowledge can be used in addition to explicit feedback. Furthermore, emerging topics can be determined. However, one of the major limitations we faced during this study, was the data availability. We only have access to a few months of data. In addition, only the interactions with the chatbot are included and not the entire customer journey on the website, e.g. it was not possible to evaluate the chatbot entirely as no information on user behaviour before and after the chat was available. It can be proposed that a tracking of users across longer periods of time and across different points of interaction is necessary to generate an integrated user model. This model should be updated with each preceding interaction of the user. Evaluating the chatbot conversations is difficult as we do not know what the user did after leaving the conversation. If a user continued to search on the website or called a customer service employee, it can be assumed that the chatbot did not provide the right answer. This information is essential to determine the impact of the chatbot's recommendations.

Moreover, access to further user information i.e. contract details or age would be valuable for user model generation. In order to develop a comprehensive user model, a complete dataset containing chatdata, clickstream data, contract details and transcriptions of phone calls with the customer service has to be generated.

A goal is to distinguish in future also between user types with different personality structures. Personality determines differences among humans in their "emotional, interpersonal, experiential, attitudinal and motivational style". Past research has found significant correlations between the personality of a user and the user's preferences in a number of domains including music, travelling and language [13][14][15]. These correlations can be successfully exploited for generating recommendations to users. User information for determining personality factors can be gathered explicitly by conducting questionnaires or implicitly using machine learning techniques [14] or even pictures [15].

Finally, it is worth noting that the proposed approach in this study is not domain-depended. In future work we will therefore also consider other domains to implement our approach and systematically evaluate our approach.

References

[1] Customer service 4.0, "Customer service 4.0 - Wie gut sind Chatbots." 2019. [Online]. Available: https://www.heise.de/downloads/18/2/5/4/1/3/4/2/Studie_chatbots.pdf

[2] Nathaniel Boisgard. State-of-the-art approaches for german language chatbot development. Diploma Thesis, TU Wien, 2018.

[3] James H Martin and Daniel Jurafsky. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Pearson/Prentice Hall, 2009.

[4] Minha Lee, Lily Frank, Femke Beute, Yvonne de Kort, and Wijnand IJsselsteijn. Bots mind the social-technical gap. In Proceedings of 15th European Conference on Computer-Supported Cooperative Work-Exploratory Papers. European Society for Socially Embedded Technologies (EUSSET), 2017.

[5] Graeme Kennedy. An introduction to corpus linguistics. Routledge, 2014.

[6] Sholom M Weiss, Nitin Indurkhya, and Tong Zhang. Fundamentals of predictive text mining. Springer, 2015.

[7] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pages 5560, 2014.

[8] Martin Rajman and Romaric Besanon. Text mining: natural language techniques and text mining applications. In Data mining and reverse engineering, pages 5064. Springer, 1998.

[9] Rensis Likert. A technique for the measurement of attitudes. Archives of psychology, 1932.

[10] I Elaine Allen and Christopher A Seaman. Likert scales and data analyses. Quality progress, 40(7):6465, 2007.

[11] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. Crisp-dm 1.0 step-by-step data mining guide. 2000.

[12] Mohamed Ramzi Haddad, Hajer Baazaoui, Djemel Ziou, and Henda Ben Ghezala. A predictive model for recurrent consumption behavior: An application on phone calls. Knowledge-Based Systems, 64:3243, 2014.

[13] David Rawlings and Vera Ciancarelli. Music preference and the five-factor model of the neo personality inventory. Psychology of Music, 25(2):120132, 1997.

[14] Marko Tkalcic and Li Chen. Personality and recommender systems. In Recommender systems handbook, pages 715739. Springer, 2015.

[15] Julia Neidhardt, Rainer Schuster, Leonhard Seyfang, and Hannes Werthner. Eliciting the users unknown preferences. In Proceedings of the 8th ACM Conference on Recommender systems, pages 309312. ACM, 2014.

[16] Mark Newman. Networks. Oxford university press, 2018.