Measuring Service Encounter Satisfaction with Customer Service Chatbots using Sentiment Analysis

Jasper Feine[1], Stefan Morana[1], and Ulrich Gnewuch[1]

[1] Karlsruhe Institute of Technology, Institute of Information Systems and Marketing (IISM), Karlsruhe, Germany
{jasper.feine,stefan.morana,ulrich.gnewuch}@kit.edu

Abstract. Chatbots are software-based systems designed to interact with humans using text-based natural language and have attracted considerable interest in online service encounters. In this context, service providers face the challenge of measuring chatbot service encounter satisfaction (CSES), as most approaches are limited to post-interaction surveys that are rarely answered and often biased. As a result, service providers cannot react quickly to service failures and dissatisfied customers. To address this challenge, we investigate the application of automated sentiment analysis methods as a proxy to measure CSES. Therefore, we first compare different sentiment analysis methods. Second, we investigate the relationship between objectively computed sentiment scores of dialogs and subjectively measured CSES values. Third, we evaluate whether this relationship also exists for utterance sequences throughout the dialog. The paper contributes by proposing and applying an automatic and objective approach to use sentiment scores as a proxy to measure CSES.

Keywords: online customer service, chatbot, sentiment analysis, service encounter satisfaction, correlation analysis

1 Introduction

Digital communication technologies have become an integral part for organizations to interact with their customers [1]. Many companies offer online services via live chat interfaces, which enable customers to directly interact with customer service employees [2]. This type of text-based service encounter is a cost effective service solution and often the preferred way of communication for young people [3]. One technology which is often deployed to assist service employees in online service encounters are chatbots [1]. Chatbots are software-based systems designed to interact with humans via text-based natural language [4, 5] and can be found across industries (e.g., airlines, energy provider). Gartner predicts that by 2020, 25% of all customer services organizations will integrate this technology [6].

Despite their great potential, many customer service chatbots did not meet customer expectations and led to service failures [7]. As a result, many service providers retired their chatbots, as unsatisfactory online service encounters have negative effects on

14th International Conference on Wirtschaftsinformatik,
February 24-27, 2019, Siegen, Germany

word-of-mouth, loyalty, and intention to repurchase a product [8]. Ignoring customer frustrations can strongly impede the performance of customer service encounters and carries the risks that the service chatbot is perceived as cold, socially indebt, untrustworthy, and incompetent [9]. Therefore, service providers should identify service encounters that were below customer's expectations [10] and trigger service recovery procedures (e.g., offering compensation). Such procedures can help to recover from almost any service failure and increase trust, perception of fairness, and service experience [10]. However, most approaches to identify dissatisfied customers in a text-based online environment (e.g., chat, social media) are limited to post-interaction surveys [11]. This is problematic as self-reported data can hardly be retrieved during an interaction, is influenced by various biases [12], and only few users are willing to provide this kind of information [8, 11]. Therefore, we propose that an automated method to measure chatbot service encounter satisfaction (CSES) during a customer-chatbot interaction could help service providers to deal with these issues.

To develop such a method, we want to take advantage of the fact that written text is associated with a person's thoughts, emotions and motivations [13–15]. Humans write differently when they are happy or frustrated and thus, written text by itself conveys much information about a human [14]. Users who are less happy with a chatbot use less assent, fewer positive, and more anger-related words and thus, express more negative sentiments [16]. The analysis of such opinionated text can provide valuable information about the user as opinions are "*key influencers of our behaviors*" [17, p. 2] and "*sentiment and tonal polarity are inherent properties of human-human communication and interaction*" [18, p. 1367].

As a manual analysis of expressed polarity in written text does not scale well to larger datasets [19], automated sentiment analysis methods have been developed. These methods are capable of automatically extracting positive or negative polarity expressed in written text [20]. Moreover, current sentiment analysis methods have been found to be very accurate and thus, seem to be a valid approach [21, 22]. However, research has rarely applied sentiment analysis in human-computer interaction (HCI) so far [20, 23]. Most HCI studies focus on auditory and visual signals of humans as these transmit the majority of communication-related information [20]. Moreover, most sentiment analysis studies focus on the method itself [23]. As a result, there is a lack of understanding on how to apply sentiment analysis in online chatbot service encounters to obtain valuable information about the user and her/his CSES. Therefore, we investigate the application of sentiment analysis methods for chatbots in online service encounters by drawing on research that text-based communication by itself is rich in informative signals [24] and that written language is influenced by emotions, intentions, and thoughts [13–15]. More specifically, we argue that sentiment analysis of dialog data can be used as an easy-to-use and objective proxy to measure CSES. Therefore, our research project addresses the following research question:

> *How to measure service encounter satisfaction with a chatbot using sentiment analysis methods?*

To address this research question, we first compare different sentiment analysis methods on an empirical level by analyzing the calculated sentiment scores for two datasets. Next, we test for a potential correlation between sentiment scores and CSES

values that were measured using a survey-based approach in an online experiment. In doing so, we first investigate this potential relationship on a dialog level and second on an utterance level (i.e., single messages). This paper contributes by proposing and applying an automatic and objective approach to use sentiment scores as a proxy to measure CSES. Our proposed approach enables researchers and practitioners, such as online customer service providers, to objectively and automatically retrieve valuable information after and during an online service encounter.

2 Related Work

2.1 Customer Service Chatbots

Recent advances in technology and great business potential have led to an increased interest in the development of conversational agents [5, 25]. Conversational agents are software-based systems designed to converse with a user via natural language [4, 5]. Thereby, the user interacts with the conversational agent in a natural dialog and does not use a predefined set of keywords or command phrases [4]. They can offer both speech- and text-based interfaces and can also be visualized and animated (i.e., embodied conversational agent) [4]. Conversational agents that interact with the user primarily via a text-based interface are often referred to as chatbots [5]. Chatbots can be deployed on various communication channels, such as instant messaging platforms (e.g., Line, Telegram, WeChat), websites, or on social media (e.g., Facebook, Twitter) and are accessible from various devices (e.g., PCs, mobile phones) [4]. Since Weizenbaum developed the first chatbot named ELIZA in 1966, much research has been conducted and various chatbots have been deployed across industries [4, 5].

One of the reasons why both research and practice are increasingly using this technology is the fact that chatbots interact in a human-like interaction style (i.e., use natural language) and offer great business potential (i.e., 24/7 availability at lost costs) [4]. Therefore, chatbots are increasingly implemented in online service encounters as many companies communicate with their customers via live-chats on their website or on social media platforms [1, 2]. Chatbots could help to automate online customer service, save costs, and enhance online experience [1, 26]. For example, instead of a customer calling or chatting with a service employee, customers are now communicating with a service chatbot [26]. In addition, chatbots can also take the role of first tier support agents and assist customer service employees. Therefore, chatbots can first start an online service encounter and then seemingly handover the conversation to a human agent when required. This can lead to a great reduction of routine requests usually handled by service employees.

2.2 Chabot Service Encounter Satisfaction

Satisfaction is an often applied construct in information systems (IS) research to evaluate the success and effectiveness of a system and it is particularly critical for the success of service systems [27]. It reflects whether customers perceive a service as

pleasurable with regard to its consumption-related fulfilment [8]. High customer satisfaction values are important to achieve long-term success, especially in highly competitive markets, and therefore should have priority for any organization [8, 28].

Customer satisfaction is strongly impacted by the service encounter satisfaction, which refers to the post-consumption evaluation of a service encounter [29, 30]. A successful service encounter makes a company's product incrementally more effective and easier to use [28], influences the customer's choice independent whether a service is provided offline or online [31], and is linked to several desired outcomes such as word-of-mouth, loyalty, and intention to repurchase a product [8, 32]. Thus, service encounter satisfaction is a critical indicator for any organization [8, 28].

Service encounter satisfaction is influenced by several antecedents such as the customization and flexibility in service encounter, effective service recovery when failures occur, and spontaneous delights (i.e., pleasing experiences customers do not expect) [32]. In addition, various design elements of a chatbot influence the CSES such as verbal communication cues (i.e., being polite, responsive, and show mutual understanding), level of expertise (i.e., a core attribute of a service employee), or visual cues (i.e., such as an avatar) [29]. In an online context, the measurement of CSES is often limited to follow-up surveys [11, 29]. Thus, CSES cannot be retrieved in real-time, is often biased, and the surveys are only answered by a few users [11, 12]. Moreover, customers have a general "*reluctance to share their sentiments with firms*" [8, p. 359] and thus, companies are often not able to react fast enough to dissatisfied customers using service recovery procedures [10]. Failing to recover can result in lost customers, negative word of mouth, decreased loyalty, and less profits [28, 32].

2.3 Sentiment Analysis Methods

A common method within the natural language understanding literature is the analysis of opinions and sentiments expressed in written text. This becomes meaningful as research has shown that written text is clearly impacted by the user's emotions, intentions, and thoughts [13–15]. Consequently, written text says something about us and can be used as a proxy for information about the author. Therefore, various methods have been developed to analyze the opinions and sentiments expressed in written text [21]. These methods are named and defined in many different ways (e.g., sentiment analysis, opinion mining, see [33]). As it is the most common name, we follow [17, 33] and define sentiment analysis as the computational analysis of written language to identify the user's perceived positive or negative valence towards a certain entity (e.g., product, service, event). Sentiment analysis has recently witnessed great attention, because of the large availability of opinion-rich resources on the Internet (e.g., online reviews) and advances in artificial intelligence [17]. Consequently, many of the major technology companies offer sentiment analysis solutions (e.g., IBM, Google) and also various open source solutions are available (see [21]). This led to the development of many available and precise methods (see [20, 21]).

Sentiment analysis methods can be generally distinguished into two broad but also overlapping approaches, namely the application of semantic rules or statistical methods [20]. Methods of the first category compare sentiment-related expressions with

sentiment lexicons that contain the semantic orientation of words [34]. One of the greatest challenges of these methods is that the semantic orientation of individual words does not necessarily correspond to the contextual polarity of the whole sentence [34]. Therefore, it is necessary to extract additional linguistic patterns of the text by conducting morpho-syntactic text analyses (i.e., wordform, lemma, part of speech tags) [20]. Too specific extraction patterns, however, limit the application range to a specific domain. Methods of the second, more recently applied category use unsupervised or supervised machine learning algorithms including support vector machines and Bayes classifiers [20]. These methods enable the development of more generic models, but require labeled data for training purposes. Consequently, the quality of such models is heavily influenced by the reliability of sentiment annotations [20].

Today's applications of sentiment analysis are manifold. Sentiment analysis can be used to predict the success of political campaigns [35], identify interaction problems within a conversation corpus [11], or even to scan the dark web in an intelligence context [36]. Nevertheless, only a few studies analyzed sentiments in a chatbot context yet as most studies are focusing on the method itself [20, 23]. One reason is the difficulty to classify rather short informal chat messages, which include a high degree of language creativity, spelling mistakes, and the expression of sentiments without real intentions [19]. Another reason are the differences and ambiguities in human mood coding which make it difficult to create a gold standard [37] and thus, it is difficult to develop user-independent prediction models [38]. However, some related research has already applied sentiment analysis to infer the customer satisfaction from product reviews for shopping websites and mobile services [23, 39].

3 Research Method

To answer our research question, we first selected suitable dialog corpora and sentiment analysis methods to run our analyses. Then, we defined a three-step research approach to analyze the corpora in order to answer our research question.

3.1 Dialog Corpora and Sentiment Methods

First, we collected one dialog corpus from an online experiment in a customer service context [40]. The participants (n = 79, mean age = 28.835, SD age = 6.388) were given a fictive mobile phone bill and the experimental task was to find a more suitable mobile phone plan through interacting with a customer service chatbot. The chatbot asked several consumption-related questions and was capable of responding interactively to given user queries. After the interaction, all participants were asked to complete a questionnaire measuring CSES using an established measurement instrument on a 7-point Likert scale [29]. The construct displayed a sufficient composite reliability (CR) above 0.8 (CR = 0.814) and the average variance extracted was above 0.5. All measurement items had factor loadings above 0.7 and the mean CSES value was 4.924 (SD = 1.179). The complete experiment, all dialogs, as well as the questionnaire were in English. The complete dialog corpus consists of 79 user dialogs with a total of 1416

user utterances. We removed 353 utterances because they consisted of only mobile contract related numbers. The final corpus included 79 dialogs and a total of 1063 utterances with an average of 13.456 utterances per dialog (SD = 8.312). We refer to this dialog corpus as "ExpCorpus" in the remainder of this paper.

In addition to ExpCorpus, we used a second, publicly available dialog corpus (without CSES values) in order to have a greater basis for the comparison of different sentiment analysis methods. Therefore, we selected the "ConvAI" dialog corpus [41]. 500 volunteers chatted with ten chatbots and the dialog set is freely available as a JSON-File. The dataset includes 2778 dialogs from which we excluded 441 human-human dialogs, 102 empty dialogs, 54 bot only dialogs, and one numbers-only dialog. This resulted in the extraction of 2180 human-chatbot dialogs, which were neither empty nor contained only numbers. Finally, we extracted all 12482 human written utterances. We refer to this dialog corpus as "ConvAI" in the remainder of this paper.

To select appropriate sentiment analysis methods for our study, we reviewed two benchmark analyses [21, 42]. We followed the benchmark analysis of [21], which compared 24 open source methods, as well as the benchmark analysis of [42], which also included sentiment analysis methods from major technology companies (e.g., IBM, Microsoft). The benchmark analyses reveal that there is no superior sentiment analysis method because all tools perform differently depending on the specific context they are applied on or depending on the corresponding data source on which they were trained [21]. Consequently, both benchmarks reveal several suitable methods depending on the respective context and the training data [21]. The benchmark of [21] reveals that two of the best sentiment analysis methods providing numerical polarity for negative, neutral, and positive sentiments are VADER [43] and AFINN (i.e., an extension of ANEW [44]) [21]. VADER and AFINN are rule-based sentiment analysis methods, which use rules and heuristics to match the analyzed texts to sentiment lexicons. Both lexicons were developed and trained on social media content and Twitter data [21, 43]. The benchmark analysis of [42] reveals that the sentiment analysis methods by IBM Watson, Google Cloud, and Microsoft Azure perform best with varying types of datasets [42]. These sentiment analysis methods leverage machine learning classification algorithms in order to predict the sentiment score. Therefore, all three providers trained their algorithms on an extensive body of sentiment annotated text databases [42]. To cover both types of sentiment analysis techniques, namely semantic rules and statistical methods [20], we selected the following methods for our study: two open source methods using rule-based sentiment analysis methods (i.e., VADER, AFINN) and three commercial methods using machine learning classification algorithms (i.e., IBM Watson, Google Cloud, and Microsoft Azure). We calculated the sentiment scores for each of the open source methods using the web service ifeel 2.0 provided by [22] and for each of the commercial methods using their Node.js APIs.

3.2 Research Approach

In this section, we present our three-step research approach (see Table 1) to answer our research question and to investigate the potential correlation between sentiments and CSES. All analyses were conducted using R 3.5.0.

Table 1. Research approach

Step	Research method	Dialog corpora	Sentiment methods
1.	Comparison of sentiment analysis methods	ConvAI (dialog & utterance level), ExpCorp (dialog & utterance level)	VADER, AFINN, IBM, Microsoft, Google
2.	Correlation analysis between sentiment scores and CSES values	ExpCorp (dialog level)	VADER, AFINN, IBM, Microsoft, Google
3.	Exploratory analysis of sentiment scores and CSES values	ExpCorp (utterance level)	IBM

In the first step, we compared all selected sentiment methods because the accuracy of sentiment analysis method are highly context and data dependent. Therefore, we investigated whether sentiment scores from each tool are similar on a dialog and utterance level by calculating the sentiment scores for each dialog and each single utterance of both corpora with all five methods. Next, we tested for potential correlations among the five sentiment scores. We do this analysis on a sentence and utterance level as sentiment analysis methods seem to perform better on "*carefully authored, lengthier content, but often struggle when faced with informal online communication*" [19, p. 318]. Consequently, we assume that some sentiment methods may struggle to predict the sentiment score of rather short utterance level and that the methods perform quite differently on both levels.

In the second step, we tested for a correlation between sentiment scores and CSES values. Therefore, we standardized the sentiment scores to -1 (i.e., negative) and +1 (i.e., positive) and subsequently tested for a correlation between sentiment scores (from all five methods) and CSES values using the dialogs and satisfaction data of ExpCorpus. By doing this, we aimed to reveal whether sentiment scores are a valid proxy for CSES values.

In the third step, we investigated the minimum number of utterances required to show a correlation between sentiment scores and CSES values. For this analysis, we used IBM's sentiment method because it yielded the highest correlation in the previous step. Therefore, we extracted utterance sequences of each dialog, calculated their sentiment scores, and tested for a correlation between sentiment scores and CSES values. Next, we investigated whether these findings also hold for utterance sequences throughout the whole dialog. This analysis provides insights whether sentiment scores can be used as a proxy for CSES during a customer service encounter.

4 Results

Step 1: Comparison of Sentiment Analysis Methods

In the first step, we started our analysis by comparing the calculated sentiment scores of selected sentiment analysis methods for both dialog corpora (ConvAI and ExpCorpus). Table 2 contains the correlation analysis between sentiment scores of both corpora for each dialog and single utterance calculated by all five methods.

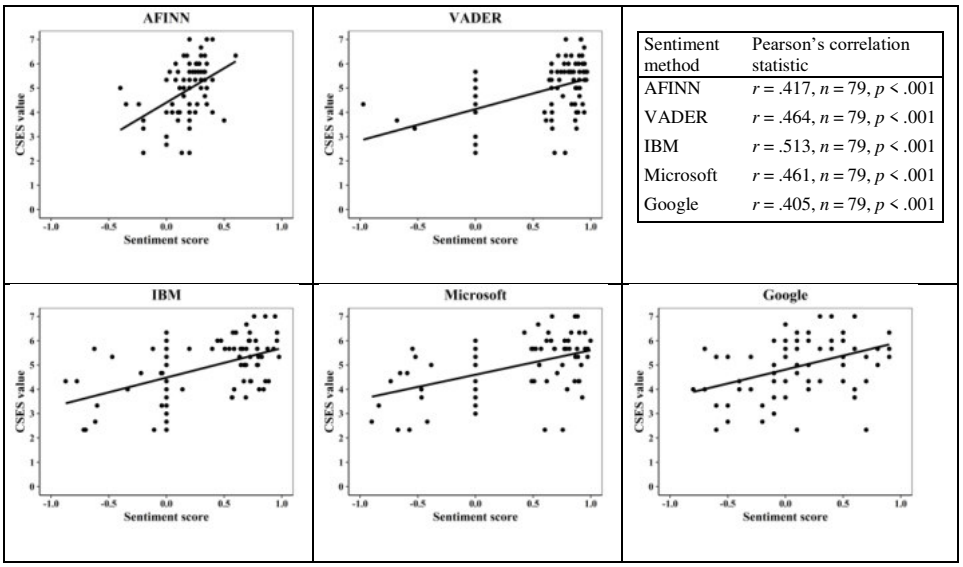
Table 2. Pearson correlation analyses among sentiment scores of different methods

Corpus	Method	AFINN	VADER	IBM	Microsoft	AFINN	VADER	IBM	Microsoft
ConvAI	AFINN	-				-			
	VADER	.605***	-			.322***	-		
	IBM	.385***	.317***	-		.387***	.357***	-	
	Microsoft	.368***	.356***	.533***	-	.300***	.311***	.604***	-
	Google	.369***	.295***	.504***	.395***	.414***	.388***	.600***	.497***
		n = 2180 dialogs				n = 12482 utterances			
ExpCorpus	AFINN	-				-			
	VADER	.719***	-			.597***	-		
	IBM	.508***	.512***	-		.505***	.169***	-	
	Microsoft	.473***	.467***	.600***	-	.383***	.201***	.615***	-
	Google	.516***	.366***	.521***	.537***	.653***	.439***	.625 ***	.487***
		n = 79 dialogs				n = 1063 utterances			
*** $p < .001$									

The results reveal that sentiment scores of dialog data from both corpora are at least moderately positively correlated with each other [45] (ConvAi $.295 \leq r \leq .605$, $n = 2180$, $p < .001$, ExpCorpus $.366 \leq r \leq .719$, $n = 79$, $p < .001$). The strongest correlation for ConvAi dialogs were identified between VADER's and AFINN's sentiment scores ($r = .605$, $n = 2180$, $p < .001$) and the weakest between Vader's and Google's sentiment scores ($r = .295$, $n = 2180$, $p < .001$). The strongest correlation for ExpCorpus was again identified between VADER's and AFINN's sentiment scores ($r = .719$, $n = 79$, $p < .001$) and the weakest one between Vader's and Google's sentiment scores ($r = .366$, $n = 2180$, $p < .001$). All sentiment scores on an utterance level were significantly positively correlated, but some correlations were weaker among some methods than they were on a dialog level ($.169 \leq r \leq .653$, $p < .001$). All in all, the findings reveal that sentiment methods using similar methodologies to identify the expressed polarity in a given text provide rather similar results. Thus, methods using semantic rules such as VADER and AFINN are strongly correlated on a dialog level. Moreover, methods using machine classification algorithms such as IBM's, Microsoft's, and Google's methods are at least moderately correlated on a dialog and utterance level.

Step 2. Correlation Analysis Between Sentiment Scores and CSES Values

In the second step, we tested for a correlation between sentiment scores and CSES values using the dialogs and CSES values of ExpCorpus. The results and the corresponding scatterplots are displayed in Figure 1. The analysis reveals a significant moderate to strong correlation between sentiment scores (from all five methods) and CSES values ($.405 \leq r \leq .513$, $n = 79$, $p < .001$) [45]. Thus, we conclude that there is a moderate positive correlation between sentiment scores and CSES values for four sentiment analysis methods and a strong positive correlation for IBM's sentiment method ($r = .513$, $n = 79$, $p < .001$) [45]. Moreover, it becomes visible that sentiment scores seem to be primarily a better predictor for positive than for negative CSES values. Moreover, semantic rule based algorithms seem to calculate sentiment scores of the dialogs generally more positive.



Sentiment method	Pearson's correlation statistic
AFINN	$r = .417, n = 79, p < .001$
VADER	$r = .464, n = 79, p < .001$
IBM	$r = .513, n = 79, p < .001$
Microsoft	$r = .461, n = 79, p < .001$
Google	$r = .405, n = 79, p < .001$

Figure 1. Correlation analyses between sentiment scores (of dialogs) and CSES values for ExpCorpus

Step 3: Exploratory Analysis of Sentiment Scores and CSES Values

In the third step, we investigated the minimum number of utterances required to show a significant positive correlation between sentiment scores and CSES values. Therefore, we combined the first ten utterances (u_i, i = 1, … ,10) into ten different utterance sequences (US_i, i = 1, …, 10), calculated their sentiment scores, and tested for a correlation with CSES values. The results are summarized in Table 3.

Table 3. Correlation analyses between sentiment scores (of utterances sequences) and CSES values for ExpCorpus

Analysed utterance squence	Included dialogs	Included utterances	Included words	Pearson' s correlation statistic
$US_1 = \{u_1\}$	79	79	739	$n = 79, r = .018, p = .872$
$US_2 = \{u_1, u_2\}$	79	158	1086	$n = 79, r = .133, p = .244$
$US_3 = \{u_1, u_2, u_3\}$	79	237	1360	$n = 79, r = .234, p = .038$
$US_4 = \{u_1, \ldots, u_4\}$	79	316	1574	$n = 79, r = .251, p = .026$
$US_5 = \{u_1, \ldots, u_5\}$	79	395	1779	$n = 79, r = .372, p < .001$
$US_6 = \{u_1, \ldots, u_6\}$	75	450	1989	$n = 75, r = .437, p < .001$
$US_7 = \{u_1, \ldots, u_7\}$	74	518	2159	$n = 74, r = .480, p < .001$
$US_8 = \{u_1, \ldots, u_8\}$	68	544	2350	$n = 68, r = .443, p < .001$
$US_9 = \{u_1, \ldots, u_9\}$	58	522	2495	$n = 58, r = .506, p < .001$
$US_{10} = \{u_1, \ldots, u_{10}\}$	46	460	2633	$n = 46, r = .503, p < .001$
All dialogs with all utterances	79	1060	3431	$n = 79, r = .513, p < .001$

Please note that not all dialogs included up to ten user utterances. As a consequence, the number of analyzed dialogs decreases with increasing sequence length. The last row analyzes all dialogs including all utterances of each dialog.

The analysis reveals that the sentiment scores of US_1 and US_2 have no significant correlation with the CSES values. However, the correlation increases with an increasing number of utterances combined in each sequence. Our results show a significant weak positive correlation between sentiment scores and CSES values after the analysis of the first three utterances ($r = .234$, $n = 79$, $p = .038$). Moreover, we revealed a significant moderate positive correlation ($r = .372$, $n = 79$, $p < .001$) after the analysis of the first five utterances [45]. To provide a greater understanding of these findings, Table 4 provides some exemplary utterance sequences, their sentiment scores, and the measured CSES values.

Table 4. Exemplary utterance sequences including the first three utterances

Utterance sequence	Sentiment score	CSES value
{"*Hi*", "*Nice to meet you. I'm interested in a cheaper phone plan. Can you help me?*", "*I think it is SuperMobile*"}	0.769	6
{"*Hey, I'm currently on the mobile phone plan Yellow Basic 1000 and I received an unexpectedly high mobile phone billl last month.*", "*Are there any better mobile phone plans for me?*", *"It's SuperMobile Yellow Basic 1000*"}	0.488	6
{"*My bill is too high*", "*Help me to find a new mobile phone plan*", "*I dont know*"}	-0.566	4,333

Having shown that at least the first three utterances of a dialog are required to find a significant positive correlation between sentiment scores and CSES values, we further investigated whether this correlation can also be found for all utterance sequences throughout the whole dialogs. Therefore, we extracted all consecutive utterance sequences consisting of three or five utterances within the first ten utterances of each dialog. This extraction resulted in eight consecutive utterance sequences for dialogs that were at least that long (e.g., *Seq-1* = {*u1*, *u2*, *u3*}, *Seq-2* = {*u2*, *u3, u4*}, *Seq-9* = {*u1*, *u2*, *u3*, *u4*, *u5*}). Then we calculated the sentiment scores and tested for a correlation between sentiment scores and CSES values (see Table 5).

Table 5. Correlation analyses between sentiment scores (of consecutive utterance sequences) and CSES values for ExpCorpus

Sequence	Included utterances	Pearson' s correlation statistic
Seq-1	{*u1, u2, u3*}	$n = 79, r = .234, p = .038$
Seq-2	{*u2, u3, u4*}	$n = 79, r = .243, p = .031$
Seq-3	{*u3, u4, u5*}	$n = 79, r = .289, p = .010$
Seq-4	{*u4, u5, u6*}	$n = 75, r = .350, p = .002$
Seq-5	{*u5, u6, u7*}	$n = 74, r = .410, p < .001$
Seq-6	{*u6, u7, u8*}	$n = 68, r = .267, p = .029$
Seq-7	{*u7, u8, u9*}	$n = 58, r = .501, p < .001$
Seq-8	{*u8, u9, u10*}	$n = 46, r = .501, p < .001$
Seq-9	{*u1, u2, u3, u4, u5*}	$n = 79, r = .372, p < .001$
Seq-10	{*u2, u3, u4, u5, u6*}	$n = 75, r = .407, p < .001$
Seq-11	{*u3, u4, u5, u6, u7*}	$n = 74, r = .436, p < .001$
Seq-12	{*u4, u5, u6, u7, u8*}	$n = 68, r = .377, p = .002$
Seq-13	{*u5, u6, u7, u8, u9*}	$n = 58, r = .503, p < .001$
Seq-14	{*u6, u7, u8, u9, u10*}	$n = 46, r = .325, p = .028$

Please note that not all dialogs included up to ten user utterances. As a consequence, the number of analyzed dialogs decreases with increasing utterance position.

The analysis shows that sentiment scores of all utterance sequences throughout the whole dialog are positively correlated with the CSES values. All correlations are significant at least at a $p < .05$ level. The correlation strength varies among the different sequences between weak and strong correlation. However, the minimum and maximum value of the correlation strength is higher for sequences consisting of five consecutive utterances, which always had at least a moderate positive correlation with CSES values.

5 Discussion

In this paper, we investigate whether sentiment scores from textual input can be used as a proxy to measure CSES in a customer-chatbot interaction. Therefore, we followed a three-step research approach: first, we compared five sentiment analysis methods by testing the relation of sentiment scores from two dialog corpora. Second, we tested for a correlation between sentiment scores and CSES values. Third, we analyzed this correlation in detail at the utterance level. Results of step 1 reveal a significant positively correlation among sentiment scores from all selected sentiment analysis methods. Results of step two reveal that sentiment scores of complete dialogs are significantly positive correlated with the subjectively measured CSES values. Results of step three reveal that this relation is not only valid for the analysis of an entire dialog, but also for any sequences of at least three consecutive utterances throughout the entire dialog. Thus, we conclude that sentiment scores can be used as an automatic and objective proxy to measure CSES in an online service encounter. Therefore, our findings further contribute to existing research that states "*sentiment analysis corresponds surprisingly well with emotional self-report*" [15, p. 87].

The results of our analysis have implications for the design of customer service chatbots. As customers may express their frustrations in written language, future chatbots could continuously perform sentiment analyses and use sentiment scores as a proxy to identify dissatisfied customers (by analyzing at least three consecutive utterances). In this way, service providers can intervene to reduce the risk of service failures. For example, a customer service chatbot could recognize that the current conversation with a customer is turning towards a negative sentiment score. In this case, several strategies could be triggered. The chatbot could seamlessly handover the conversation to a trained human service agent, automatically trigger service recovery procedures, or express certain verbal utterances such as excuses [46, 47]. Research has shown that these immediate reactions can reduce the level of frustration [46] and can lead to an increased interaction length [47]. Furthermore, service providers can use this data in post-interaction analyses to retrieve valuable information about CSES. This information cannot only be used for service recovery, but also for identifying general weaknesses in the service quality of the chatbot.

Although we aimed to ensure a high rigor in our research, some limitations should be considered. First, many sentiment analysis methods exist and they all may evaluate a given text differently depending on the context and type of a message [21]. This becomes even more meaningful when applied to rather short and informal chat data. Therefore, a different selection of sentiment methods may have led to different results.

Consequently, we tried to minimize this risk by starting with a selection of five sentiment methods based on benchmarks and compared them with each other by applying them on two dialog corpora. Even though all sentiment analysis methods had a moderate to strong correlation to CSES, some sentiments methods were rather weak predictors for users having low CSES values. Therefore, it "*is important that researchers and companies perform experiments with different methods before applying a method*" [21, p. 27]. Second, we analyzed a dialog corpus, which measured the CSES using a post-interaction survey. However, data of a survey-based approach might be influenced by various biases [12]. To reduce this risk, we reviewed all dialogs and verified that participants followed the experimental task and did not answer with straight line responses. Third, we only analyzed the relationship between sentiments and CSES based on a dialog corpus from a hypothetical online service task (i.e., finding new plan) in a specific context (i.e., mobile contract) in one language (i.e., English). Therefore, it is unclear whether our findings also hold for other customer service tasks (e.g., book ticket) in other contexts (e.g., airlines) in other languages (e.g., German). Fourth, we conducted correlation analyses between sentiment scores and CSES values to reveal a correlation between these two variables. Even though we found a strong positive correlation and propose sentiment scores as a proxy for CSES values, this analysis does not provide the explanation for this relation and does not indicate a cause-and-effect relationship [48]. Thus, results need to be applied with care as we cannot predict CSES based on sentiments scores or vice versa [48].

Considering these limitations, we identify several avenues for future research. First, future work can replicate our analyses on additional dialog corpora from different contexts, doing different tasks, and in different languages. This could further strengthen the applicability of sentiment analysis as a proxy to measure CSES in several domains and languages. Second, future studies could investigate adaptive reaction strategies based on real-time analyses of at least three consecutive user utterances. This could enable chatbots to recognize user frustrations and supports the development of chatbots that act more socially [46, 47]. Moreover, future research could investigate the application of more trivial text analysis methods, such as word count and length of dialogs, as well as more complex methods, such as topic modelling, as proxies to predict customer satisfaction. Integrating these techniques into a chatbot can lead to even greater understanding of the user and enables more precise reactions by the chatbot.

6 Conclusion

In this paper, we investigate the application of sentiment analysis methods in an online service encounter with a chatbot and show that sentiment scores can serve as a proxy to measure CSES. This enables researchers and practitioners, such as online service providers, to objectively and automatically retrieve user information during and after an online service encounter. This information can be used not only to trigger service recovery procedures, but also to identify weaknesses in the service quality and to analyze the user in real-time. Therefore, our results contribute towards the design of user adaptive service chatbots.

References

1. Larivière, B., Bowen, D., Andreassen, T.W., Kunz, W., Sirianni, N.J., Voss, C., Wünderlich, N.V., Keyser, A. de: "Service Encounter 2.0": An investigation into the roles of technology, employees and customers. Journal of Business Research 79, 238–246 (2017)
2. McLean, G., Osei-Frimpong, K.: Examining satisfaction with the experience during a live chat service encounter-implications for website providers. Computers in Human Behavior 76, 494–508 (2017)
3. Kowatsch, T., Nißen, M., Rüegger, D., Stieger, M., Flückiger, C., Allemand, M., Wangenheim, F. von: The Impact of Interpersonal Closeness Cues in Text-based Healthcare Chatbots on Attachment Bond and the Desire to Continue Interacting: An Experimental Design. In: Twenty-Sixth European Conference on Information Systems (ECIS). Portsmouth, UK (2018)
4. McTear, M., Callejas, Z., Griol, D.: The Conversational Interface. Talking to Smart Devices. Springer International Publishing, Switzerland (2016)
5. Dale, R.: The return of the chatbots. Natural Language Engineering 22, 811–817 (2016)
6. Gartner: Gartner Says 25 Percent of Customer Service Operations Will Use Virtual Customer Assistants by 2020, https://www.gartner.com/newsroom/id/3858564
7. Ben Mimoun, M.S., Poncin, I., Garnier, M.: Case study—Embodied virtual agents. An analysis on reasons for failure. Journal of Retailing and Consumer Services 19, 605–612 (2012)
8. Oliver, R.L.: Satisfaction. A behavioral perspective on the consumer. McGraw Hill, New York (1997)
9. Brave, S., Nass, C.: Emotion in human-computer interaction. In: Jacko, J.A., Sears, A. (eds.) The human-computer interaction handbook, pp. 81–96. L. Erlbaum Associates Inc, NJ, USA (2002)
10. Holloway, B.B., Beatty, S.E.: Service Failure in Online Retailing: A Recovery Opportunity. Journal of Service Research 6, 92–105 (2003)
11. Xiang, Y., Zhang, Y., Zhou, X., Wang, X., Qin, Y.: Problematic situation analysis and automatic recognition for chinese online conversational system. In: Joint Conference on Chinese Language Processing. Wuhan, China (2014)
12. Podsakoff, P.M., MacKenzie, S.B., Lee, J.-Y., Podsakoff, N.P.: Common method biases in behavioral research: a critical review of the literature and recommended remedies. The Journal of applied psychology 88, 879–903 (2003)
13. Tausczik, Y.R., Pennebaker, J.W.: The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. Journal of Language and Social Psychology 29, 24–54 (2010)
14. Nerbonne, J.: The Secret Life of Pronouns. What Our Words Say About Us. Literary and Linguistic Computing 29, 139–142 (2014)
15. Küster, D., Kappas, A.: Measuring Emotions Online: Expression and Physiology. In: Holyst, J.A. (ed.) Cyberemotions: Collective Emotions in Cyberspace, pp. 71–93. Springer International Publishing, Cham (2017)
16. Skowron, M., Rank, S., Theunis, M., Sienkiewicz, J.: The Good, the Bad and the Neutral: Affective Profile in Dialog System-User Communication. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) Affective Computing and Intelligent Interaction, pp. 337–346. Springer, Berlin, Heidelberg (2011)
17. Liu, B.: Sentiment analysis and opinion mining. Synthesis lectures on human language technologies 5, 1–167 (2012)

18. Banchs, R.E.: On the construction of more human-like chatbots: Affect and emotion analysis of movie dialogue data. In: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Kuala Lumpur (2017)
19. Brooks, M., Kuksenok, K., Torkildson, M.K., Perry, D., Robinson, J.J., Scott, T.J., Anicello, O., Zukowski, A., Harris, P., Aragon, C.R.: Statistical Affect Detection in Collaborative Chat. In: Conference on Computer Supported Cooperative Work, pp. 317–328. ACM, New York, NY, USA (2013)
20. Clavel, C., Callejas Z.: Sentiment Analysis: From Opinion Mining to Human-Agent Interaction. IEEE Transactions on affective computing 7, 74–93 (2016)
21. Ribeiro, F.N., Araújo, M., Gonçalves, P., André Gonçalves, M., Benevenuto, F.: SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Science 5, 23 (2016)
22. Diniz, J.P., Bastos, L., Soares, E., Ferreira, M., Ribeiro, F., Benevenuto, F.: ifeel 2.0: A multilingual benchmarking system for sentence-level sentiment analysis. In: 10th international AAAI conference on weblogs and social media. Cologne, Germany (2016)
23. Kang, D., Park, Y.: Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach. Expert Systems with Applications 41, 1041–1050 (2014)
24. Walther, J.B., Parks, M.R.: Cues filtered out, cues filtered in. In: Knapp, M.L., Daly, J.A. (eds.) Handbook of Interpersonal Communication, pp. 529–563. SAGE, Thousand Oaks, CA, USA (2002)
25. Maedche, A., Morana, S., Schacht, S., Werth, D., Krumeich, J.: Advanced User Assistance Systems. Business & Information Systems Engineering 58, 367–370 (2016)
26. Gnewuch, U., Morana, S., Maedche, A.: Towards Designing Cooperative and Social Conversational Agents for Customer Service. In: Proceedings of the 38th International Conference on Information Systems (ICIS). AISel, Seoul (2017)
27. Au, N., Ngai, E.W.T., Cheng, T.E.: A critical review of end-user information system satisfaction research and a new research framework. Omega-International Journal of Management Science 30, 451–478 (2002)
28. Jones, T.O., Sasser, W.E.: Why satisfied customers defect. Harvard Business Review 73, 88-& (1995)
29. Verhagen, T., van Nes, J., Feldberg, F., van Dolen, W.: Virtual Customer Service Agents. Using Social Presence and Personalization to Shape Online Service Encounters. Journal of Computer-Mediated Communication 19, 529–545 (2014)
30. Caruana, A.: Service loyalty. European Journal of Marketing 36, 811–828 (2002)
31. Shankar, V., Smith, A.K., Rangaswamy, A.: Customer satisfaction and loyalty in online and offline environments. International Journal of Research in Marketing 20, 153–175 (2003)
32. Bitner, M.J., Brown, S.W., Meuter, M.L.: Technology Infusion in Service Encounters. Journal of the Academy of Marketing Science 28, 138–149 (2000)
33. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2, 1–135 (2008)
34. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. Computational Linguistics 35, 399–433 (2009)
35. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with twitter: What 140 characters reveal about political sentiment. In: International AAAI Conference on Weblogs and Social Media, 10, pp. 178–185. Menlo Park, CA, USA (2010)
36. Abbasi, A., Chen, H.: Affect Intensity Analysis of Dark Web Forums. In: Proceedings of Intelligence and Security Informatics. Chengdu, China (2017)

37. Thelwall, M., Buckley, K., Paltoglou, G., Di Cai, Kappas, A.: Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology 61, 2544–2558 (2010)
38. Higashinaka, R., Minami, Y., Dohsaka, K., Meguro, T.: Issues in Predicting User Satisfaction Transitions in Dialogues: Individual Differences, Evaluation Criteria, and Prediction Models. In: Lee, G.G., Mariani, J., Nakamura, S. (eds.) Spoken Dialogue Systems for Ambient Environments. Seond International Workshop, IWSDS 2010, Gotemba, Shizuoka, Japan, October 1-2, 2010. Proceedings, 6392, pp. 48–60. Springer, New York (2010)
39. Wang, Y., Lu, X., Tan, Y.: Impact of product attributes on customer satisfaction: An analysis of online reviews for washing machines. Electronic Commerce Research and Applications 29, 1–11 (2018)
40. Gnewuch, U., Morana, S., Adam, M., Maedche, A.: Faster Is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction. In: in Proceedings of the 26th European Conference on Information Systems (ECIS), Portsmouth, United Kingdom, June 23-28.
41. Logacheva, V., Burtsev, M., Malykh, V., Poluliakh, V., Rudnicky, A., Serban, I., Lowe, R., Prabhumoye, S., Black, A.W. and Bengio, Y.: A Dataset of Topic-Oriented Human-to-Chatbot Dialogues, http://convai.io/2017/data/dataset_description.pdf (Accessed: 30.08.2018)
42. Corredera Arbide, A., Romero, M., Moya Fernández, J.M.: Affective computing for smart operations: a survey and comparative analysis of the available tools, libraries and web services. International Journal of Innovative and Applied Research 5, 12–35 (2017)
43. Gilbert, C.H.E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Eighth International AAAI Conference on Weblogs and Social Media. Ann Arbor, MI, USA (2014)
44. Nielsen, F.Å.: A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. arXiv preprint arXiv:1103.2903 (2011)
45. Cohen, J.: A power primer. Psychological bulletin 112, 155–159 (1992)
46. Hone, K.: Empathic agents to reduce user frustration. The effects of varying agent characteristics. Interacting with Computers 18, 227–245 (2006)
47. Klein, J., Moon, Y., Picard, R.W.: This computer responds to user frustration. Theory, design, and results. Interacting with Computers 14, 119–140 (2002)
48. Taylor, R.: Interpretation of the Correlation Coefficient: A Basic Review. Journal of Diagnostic Medical Sonography 6, 35–39 (1990)