Customer Sentiment in Web-Based Service Interactions: Automated Analyses and New Insights

Galit B. Yom-Tov
Technion—Israel Institute of Technology
Haifa, Israel
gality@technion.ac.il

Shelly Ashtar
Technion—Israel Institute of Technology
Haifa, Israel
shellya@campus.technion.ac.il

Daniel Altman
Technion—Israel Institute of Technology
Haifa, Israel
altmand@campus.technion.ac.il

Michael Natapov
LivePerson Inc.
Tel-Aviv, Israel
michaelna@liveperson.com

Neta Barkay
LivePerson Inc.
Tel-Aviv, Israel
netabarkay@gmail.com

Monika Westphal
Technion—Israel Institute of Technology
Haifa, Israel
westphal@campus.technion.ac.il

Anat Rafaeli
Technion—Israel Institute of Technology
Haifa, Israel
anatr@technion.ac.il

ABSTRACT

We adjust sentiment analysis techniques to automatically detect customer emotion in on-line service interactions of multiple business domains. Then we use the adjusted sentiment analysis tool to report insights into the dynamics of emotion in on-line service chats, using a large dataset of telecommunications customer service interactions. Our analyses show customer emotions start out negative and evolve into positive feelings, as the interaction unfolds. Also, we identify a close relationship between customer emotion dynamics *during* the service interaction and the concepts of service failure and recovery. This connection manifests itself in customer service quality evaluations *after* the interaction ends. Our study highlights the connection between customer emotion and service quality as service interactions unfold, and suggests the use of sentiment analysis tools for real-time monitoring and control of web-based service quality.

KEYWORDS

Customer service; sentiment analysis; customer satisfaction

ACM Reference Format:
Galit B. Yom-Tov, Shelly Ashtar, Daniel Altman, Michael Natapov, Neta Barkay, Monika Westphal, and Anat Rafaeli. 2018. Customer Sentiment in Web-Based Service Interactions: Automated Analyses and New Insights. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3184558.3191628

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191628

1 INTRODUCTION

The service industry is undergoing a digital revolution. Services become more and more automatic and easy to use, and service companies become more accessible through new service channels and social media (e.g. Twitter or Facebook), corporate websites, or messaging applications (e.g. WhatsApp). Still, people find service very frustrating and emotionally demanding. Available theory clearly indicates that customer service interactions envelop multiple manifestations of emotions (cf., Affective Events Theory [32]), and that emotion dynamics are important to recognize because they reflect service quality [18]. In addition, providing service through digital interfaces opens new opportunities to explore human behavior in service systems that were not available in the past [23]. Therefore, understanding the effects and dynamics of emotions that customers express, is critical.

In this study, we focus on a large on-line telecommunications company whose customers seek service through textual platforms. We aim to understand the effect of changes in communicated sentiment through the service interaction. We leverage automated sentiment analysis to analyze emotions in the service chats of this company; but instead of examining emotion in an entire interaction (as done in analyses of customer reviews), we examine changes of sentiment from a longitudinal standpoint. (Such use of sentiment analysis was done for example, in the context of health-care informatics [34] to detect progression of patient emotions.) We seek to answer the following three research questions: (1) how does customer sentiment change within a service interaction; (2) is there a connection between such changes and service quality measures; and (3) does emotion in different stages of an interaction connect to different stages in a service process, such as service failure and recovery?

We find that available sentiment analysis tools have limited accuracy when applied to detecting emotion in customer service

interactions. Research acknowledged that sentiment tools should be adjusted to the context studies (e.g. [34]). We therefore began our study by building a sentiment analysis tool adjusted to the context of customer service and validating it using a dataset of chat services of multiple domains. This includes three adjustments: (1) adjustments to the domain of customer service; (2) adjustments to specific features of specific brands; (3) adjustments to specific language features used by service customers. We then use the tool to test theoretically derived hypotheses about the dynamics of customer emotion in service interactions. Our findings offer new insights into how emotions that customers express relate to the effectiveness and quality of the service interaction.

1.1 Nature of Chat-Based Service Interactions

Customer-service chat interactions comprise a sequence of interdependent messages between customers and service employees (see Figure 1). Chat service interactions can be viewed on two levels: (i) the atomic level of individual messages, implying identification of the emotion in each individual customer message; (ii) the cumulative level of full interactions, implying identification of emotion of a complete interaction. Identifying emotion at the individual message level enables real-time detection of a customer's emotional state at the specific point in time of this particular message; an emotion score at the full interaction level provides far less granularity, but is the current industry norm, and considered an indication of overall service quality. We suggest here that analyses at the individual message level is the right way to obtain real-time assessments of overall service quality. A look limited to the full service interaction level misses meaningful distinctions between the (initial) service failure stage and potential progression toward service recovery.

Another characteristic of customer chat texts is their spontaneous and unedited language; they typically comprise short sentences, do not necessarily maintain coherence or grammatical structure, and often include shortcuts, slang, typos and spelling mistakes. Text-based interactions can also contain obscenities and extensive use of punctuation, symbols, emoticons and capitalization; these may relate to emotions of the writer. This is different from product reviews—commonly used for developing and testing sentiment analysis engines—that typically go through substantial editing, and include well thought out and socially polite text. Recent research on sentiment in Twitter takes some of these features into account (see for example [1]) but, to our knowledge, previous work only examined specific parts of an interaction and did not examine emotion dynamics that occur throughout whole customer service interactions [12]. Thus, our paper suggests that available models for automated emotion detection need to be adjusted to the context of spontaneous, real-life, text-based customer service interactions. We fill this gap by by providing a tool with specific features that fit the bill, and show insights into emotions expressed by customers interacting through chat with service employees.

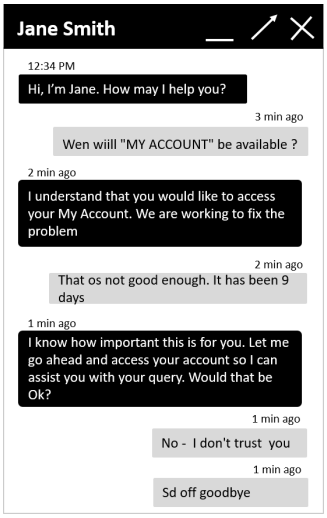

Figure 1: An example of a service interaction between employee and customer through chat

2 BACKGROUND AND HYPOTHESES DEVELOPMENT

2.1 Customer Emotion Dynamics within Service Interactions

We claim that customer emotion during service interactions is ***dynamic*** (rather than stable or constant). Following the logic of customers approaching a service provider because of a "service failure" [18, 19], we expect customer emotions when an interaction starts to be relatively negative or neutral; customers request service when they have a problem, which brings about negative emotions [11]. Customers may also express negative emotions because they believe it will get them better results [24, 30], and/or decrease service time [20]. Negative emotions early in the interaction may also occur just because people have to spend time and effort on something they feel should not have happened (i.e., a service failure). Customer negative emotions will be evident in customer expressions such as (e.g. "I need to cancel my cellular plan") or a problem (e.g. "My phone connection doesn't work!"). The implicit logic is something like: "It is the company's fault that I need to waste my time for this service". The psychology theory of Cognitive Appraisal [9] suggests that such perceptions of the need to turn to a service agent damage people's sense of well-being, and evoke negative emotions [8]. The role of service delivery agents, in turn, is to resolve problems that customers raise, and promote "service recovery" [18, 19]. Service recovery in an interaction may be more or less effective, depending on multiple factors [8]. Regardless of these factors, however, the effectiveness of the service recovery is likely to manifest in the change of the emotions that customers feel and express. Thus, we propose that customer emotion is not a stable state throughout service interactions; rather customer emotions are dynamic, and evolve during the interactions depending on the degree to which their needs are addressed and their problem is solved. This is the first unique analysis that our approach allows,

and our first prediction:

Hypothesis 1. *Customer emotions during service interactions are dynamic, evolving from initial relatively negative (based on a service failure logic), into more positive emotions at the end of the interaction (based on a service recovery logic).*

To test Hypothesis 1, we will assess customer emotion in different parts of a large sample of service interactions. Our analysis depicts the emotions that customers express as they progress through a service interaction. Obviously, different customers bring different needs, problems, and expectations, so emotion in individual interactions are likely to vary. We will depict typical emotion expressions, by reporting the average sentiment expressed in similar points in time of different interactions.

2.2 Relating Customer Emotion to Service Quality

Hypothesis 1 refers to two important parts of service interactions: the beginning and the end. Our next hypothesis regards the meaning of the dynamics between these two points. We specifically suggest that the difference in emotions between them connects to customers' overall assessment of the complete service interaction, and of the extent to which the interaction effectively resolved their problem. If the initial (relatively negative) emotion does not change during the interaction, this means the customer issue has not been resolved, and subsequently customers are *less likely* to rate the interaction as satisfying and effective, than if the change in emotion during the interaction is substantial. Thus, we predict that differences in customer emotions dynamics in successful vs. unsuccessful service interactions are meaningful indicators of service quality:

Hypothesis 2. *The magnitude of change in customer emotion during a service interaction from negative (in the start) to positive (at the end) reflects the quality of the service a customer received.*

Hypothesis 2 can refer to two aspects of customer assessments of service quality: The extent to which the customer problem was resolved in the interaction, and the extent to which a customer was satisfied with the interaction. This hypothesis is important because it suggests that customer perceptions of service quality can be assessed *during a service interaction*, rather than *after the service*, as typically done today. It also suggests that one should not simply bundle the emotions throughout the interaction together, as the emotions at the beginning of an interaction should serve as a base-line for the level of service failure the customer started with, while the trajectory within the service interaction should serve as a measure for success in service recovery. Our analyses specifically show that quantifying and dynamically assessing customer emotion *within* service interactions can predict, and potentially replace, measures of service performance *after* the interaction. We test and support the hypothesis with two popular measures of service performance, both currently collected from customers after their service interactions ended: (i) Problem resolution (known in the service industry as ***FCR***, which stands for *First Contact Resolution*), (ii) Customer satisfaction (known in the service industry as ***CSAT***). Thus, we suggest a novel way to assess service quality, using objective, unobtrusive analyses of customer expressions during an interaction.

3 METHODS

Our paper has two methods parts: Part 1 presents the sentiment analysis tool for service interactions we call ***CustSent***, and its validation. Part 2 describes insights about customer emotions and tests hypotheses using this tool.

3.1 Part 1: CustSent—A Sentiment Analysis Tool Adjusted for Service Interactions

We developed a lexicon-based model, because this approach allowed us to adapt CustSent to different service domains and brands; airline, telecommunications, or financial services share a focus on service, but may vary in specific lexicon. The alternative, machine learning approach would require training a separate model for each service domain, which would imply an extremely demanding annotation cost with each new context. See [27] for more discussion on lexicon based vs. machine learning approaches.

The model assigns an emotion score to each customer message by applying a set of rules. The score is assigned at the semantic level of a sentence. Each rule assigns a numeric integer score to words or nonverbal elements of the sentence. Each sentence is scored by multiple rules, and the set of scores is aggregated into an overall emotion score assigned to each sentence. If a message contains more than one sentence the emotion scores of the sentences are added up. A total message score above zero means total emotion of the message is *positive*; scores below zero indicate total emotion of the message is *negative*. A score of zero indicates *no emotion*[1].

Two types of rules determine the sentence emotion score: Lexicon rules assign a ***base score*** to emotionally charged words (***anchors***); anchors are manually annotated words that compose lexicons of different base polarity and intensity; e.g., positive words: *excellent (+2), great (+1), like (0)*, and negative words: *horrible (-2), confused (-1)*. The lexicons were derived inductively by looking through a large collection of customer interaction data. These lexicons are of different sizes and overall comprise a few thousands of anchors. Comparing to the available sentiment word lists (e.g. the well known Bing Liu sentiment corpus) they contain adjustments of three types:

- **Service-domain related adjustments:** We exclude or add to the sentiment lexicons words due to their special use in the service domain context. E.g., exclude words like *support, confirm, approve*; include words like *cancel, legal, waiting, elsewhere*.

  I'll take a *legal* action
  I gonna look *elsewhere* (if you don't suit me here)

  Some words even changed polarities: *promises* which seems positive, has negative connotation in service:

  I'm tired of your *promises*.
- **Business-domain related adjustments:** We exclude or add words to the sentiment lexicons due to their special use in a specific brand context, or general business context.

[1] A value of zero may also indicate an equal amount of positive and negative emotion in the same message, but our data show this occurs in a negligible number of messages.

E.g., exclude words like *gold, advanced, enhanced, premium, free, secure, solid, unlimited; miss, missed, limited, complex, blind, fall, dark, split, cold.*

Premium account, *Advanced* Program

- **Customer language adjustments:** Exclude words like *well, right, ok.* Include common misspellings of emotionally charged words, slang and obscenities in the corresponding lexicons.

Beyond the lexicons, the rules account for the ***context*** of the anchor, which is defined as the presence of negation and/or intensification words in three words preceding an anchor[2]. An anchor that appears without negation and/or intensification is considered as ***without a context***, and its base score remains unaltered. The model sums up all the context-defined scores of anchors in each sentence, creating the preliminary emotion score of the sentence.

The second set of rules updates the preliminary score of the sentence, based on features which do not change their meaning in presence of intensification and/or negation, but project an emotional charge of the whole sentence. These features also reflect the customer service and brand related context and include non-verbal (exclamation or question marks and emoticons), and verbal terms (e.g. *sorry, thanks*, or *lol*—an acronym for *laughing out loud*).

3.1.1 Lexicon Based Rules. We use five lexicons with different levels of base sentiment polarity: ***negative*** (base score -1), ***very negative*** (-2), ***positive*** (+1), ***very positive*** (+2) and ***tentative positive*** (base score 0, and becomes negative if negated).

For each lexicon context, the adjustment rule shifts the base score in cases of intensification and negation. The first four lexicons—negative, very negative, positive, very positive—follow similar adjustment rules:

- ***Intensification*** words amplify the base score of an anchor by 1 point:

pleased (+1) → *very pleased* (+2)
disappointed (-2) → *extremely disappointed* (-3)

- ***Negation*** words shift the base polarity of an anchor by 2 points in the direction of the opposite polarity (cf. [27]):

pleased (+1) → *not pleased* (-1)
disappointed (-2) → *not disappointed* (0)

- These two rules are applied in the same manner when combined:

not pleased (-1) → *very not pleased* (-2)
extremely disappointed (-3) → *not extremely disappointed* (-1)

The tentative positive lexicon is different from the other four. It comprises words (e.g. *enough, like, support, efficient, good*) which may convey positive emotion in certain cases, but in customer service interactions are used differently. Consider for example the word *like*. Most (>90%) of the appearances of the word *like* without a context have no positive connotation: most common use of the no-context *like* is neutral *"I would like to..."*. In contrast, negation of the word *like* (as in *"I don't like"*) almost always has a negative connotation. To account for such behavior we include terms such as *like* in the tentative positive category, i.e. model it as neutral without context, and negative with negation:

like (0) → *don't like* (-1) → *really don't like* (-2)

3.1.2 Sentence Level Rules.

- **Question rule**: A question structure has a different emotional load than a declarative sentence with the same wording [13, 14, 33], because questions reduce the intensity of the emotion that an anchor term expresses. For example, compare the following sentences:

I want to return it because *I don't like it.* (-1)
What is the return policy in case *I don't like it*? (0)

- **Politeness and Condition rules**: Specific verbal features, like polite words (e.g. *sorry, apologize*) or condition words (e.g. *if, maybe*), do not have a polarity score on their own, but serve as modifiers of the emotion a sentence conveys. Specifically, the model reduces the intensity of the emotion score of a sentence when politeness and/or conditioning are present:

I am *confused...* (-1) → *Sorry*, I am *confused...* (0)

- **Positive slang**: Phrases such as *yes*, *lol!*, and *no, lol!*, indicate emotionally similar (very positive in our model) reactions to an employee suggestion. A sentence sentiment score is increased in presence of such slang words.
- **Emoticons**: A check of frequencies showed that emoticons used were almost solely smilies, e.g. *:-)* and frownies, e.g. *:(*, and we consider them as non verbal indicators of emotions. They add to or subtract from the sentence score, respectively.
- **Negative idioms**: Some stable phrases and idioms—*been waiting*, *fed up*, or *your fault*—implicitly convey emotion because of the associations they insinuate. Such phrases subtract from the sentence sentiment score:

I've *been waiting* on line for over an hour now (-2)

- **Thank-you phrases**: Phrases conveying customer thanks add a positive factor to the sentiment score of a sentence in which they appear. The positive factor depends on the degree of the conveyed *thanks*, e.g.:

no, thanks (+1)
thank you sooo much for your help! (+3)

- **Multiple punctuation**: A common expression that appears in customer messages is multiple exclamation and/or question marks. Inductive analysis led us to model several patterns for such expressions. A preliminary sentiment score may be increased or decreased by multiple punctuation, e.g.:

great (+1) → *great!!!* (+2)
hello (0) → *hello???* (-2)

More sentence level rules, e.g. special attention to CAPITALIZATION patterns, were tested and rejected as not improving the model accuracy.

3.2 Assessing the Accuracy of the CustSent Model

A sample of 600 customer messages was manually annotated by three annotators (see below). To ensure coherency, we provided guidelines and examples to the annotators. We discussed dilemmas about coding, until there was agreement about the emotion in a text

[2]We compared a model with 2, 3, 4 and 5 preceding words and found 3 words to be optimal in identifying emotion in interactions conducted in English.

(ICC =.89); thus, coding was done by multiple judges, supported by consensus resolution [3, 7, 15, 21].

An initial, pilot-phase of coding (of a different sample of 200 messages) showed a majority (~70%) of messages as containing no emotion, with CustSent detecting even less emotion. Therefore, we used a stratified approach of sampling customer messages for the validation corpora. This sample is not merely a random set of messages, as this would generate a large subset of no-emotion messages. The sample includes a lower proportion of no-emotion messages than a random sample. Specifically, we considered customer messages from service chats conducted in two service brands (telecommunications and retail) during the first week in March 2016. We divided the messages into three emotion polarity groups detected by CustSent (negative, positive, no emotion), to which we refer as *negative*, *positive* and *neutral stratum*, respectively. Then we sampled an equal number of messages from each stratum. We aimed for a sample of 600 messages – 200 from each stratum. Due to technical issues, the human coders coded an effective sample comprising 597 customer messages.

Use of a sample (as opposed to a predefined Golden Standard) requires an adjustment of formulas for the accuracy metrics. We now show how the precision and recall of a sentiment analysis tool on negative emotion class is evaluated. Precision and recall for positive emotion are similarly adjusted. To measure precision and recall of negative emotion class, one compares the number of messages detected as negative by a sentiment detection tool to the number of messages coded as negative by human judges. ***Precision*** is the proportion of correct detections, and ***recall*** is the proportion of real negative emotions that are detected [16]. Formally, we denote α_{neg} the number of messages detected as negative by the sentiment analysis tool, β_{neg} the number of messages coded as negative by human judges, and γ_{neg} the number of messages detected as negative by the tool and coded as negative by human judges. Then

$$Precision(negative) = \frac{\gamma_{neg}}{\alpha_{neg}} \tag{1}$$

$$Recall(negative) = \frac{\gamma_{neg}}{\beta_{neg}} \tag{2}$$

Now, we adjust Formulas (1) and (2) by assigning to each message a weight, which is equal to the proportion of the stratum in the population it represents. Formally, let N_1, N_2, and N_3 denote the size of the negative, positive and neutral stratum, respectively. Then, a message from the i-th stratum has the weight $w_i = N_i/(N_1 + N_2 + N_3)$. Each message coded as negative by the human judges contributes its weight to the precision and recall formulas. Denote α_i^M as the number of messages detected as negative by model M in stratum i, β_i as the number of messages coded as negative by human judges in stratum i, and γ_i^M as the number of messages detected as negative by model M and coded as negative by human judges in stratum i. Hence, the precision and recall of identifying negative emotion by the model M is now:

$$Precision_M(negative) = \frac{\sum_{i=1}^{3} \gamma_i^M \times w_i}{\sum_{i=1}^{3} \alpha_i^M \times w_i} \tag{3}$$

	Negative emotion class			
Model	*Precision*	*Recall*	F_1	$F_{0.5}$
CustSent	**0.719**	0.236	0.355	**0.51**
Stanford	0.335	**0.509**	**0.404**	0.36
LIWC	0.479	0.115	0.186	0.294
SentiStrength	0.494	0.216	0.3	0.393

Table 1: Comparing four models in detecting negative emotion in customer messages.

$$Recall_M(negative) = \frac{\sum_{i=1}^{3} \gamma_i^M \times w_i}{\sum_{i=1}^{3} \beta_i \times w_i} \tag{4}$$

Note, that since all the messages detected as negative by CustSent belong to the first stratum, Formula (3) for $Precision_{CustSent}$ is the same as Formula (1).

In addition to precision and recall, we also report $\mathbf{F_1}$—the harmonic average of precision and recall—a standard way to aggregate these two into one metric.

Also, we would like to promote the use of sentiment tools for real-time assessment of customer sentiment. Such use must minimize false alarms (inaccurate alerts of negative emotion) and avoid overoptimistic inaccurate reports of positive emotion. We therefore put more emphasis on precision, especially that of negative emotion, as one of the key accuracy metric for our assessment of customer sentiment. To this end we deploy the $\mathbf{F_{0.5}}$ metric, a variation of F_1, in which precision is weighed twice as important as recall [16].

All these metrics—precision, recall, F_1, $F_{0.5}$—are calculated separately for the negative and positive emotion classes for CustSent, Stanford Sentiment Analysis RNTN model [26], SentiStrength [29], and LIWC [28], as summarized in Tables 1 and 2.[3]

CustSent outperforms previously available automatic detection models in the precision of detecting negative emotion; its precision level is significantly higher than the other models (Table 1; $p < 0.001$[4]). In recall CustSent falls behind the Stanford engine, but the precision of the latter is extremely low, and thus the $F_{0.5}$ value of CustSent is the highest among the compared detection models.

In assessments of positive emotions, CustSent has better precision than other models, though comparable to SentiStrength ($p = 0.149$). In recall, CustSent falls behind other engines ($p < 0.03$), and the $F_{0.5}$ of CustSent is similar to SentiStrength (Table 2). All in all, we show that CustSent provides the most valid customer emotion detection in service interactions, and its performance is superior to that of the other models.

3.3 Part 2: Data—Using an Automated Engine to Assess Customer Emotion in Service Chats

We used CustSent to analyze customer emotions in service chats of companies in several domains. Results are robust across domains.

[3] We first calculated the metrics separately for messages from the different firms. The results were not substantially different. For lack of space, we present the metrics of the combined sample of 597 messages, where the weight of each message corresponds to its proportion in the population.

[4] P-values reported in this section refer to a comparison of CustSent to the best result in the same category.

	Positive emotion class			
Model	*Precision*	*Recall*	F_1	$F_{0.5}$
CustSent	**0.866**	0.569	0.687	**0.784**
Stanford	0.546	0.339	0.418	0.486
LIWC	0.491	**0.717**	0.583	0.524
SentiStrength	0.813	0.677	**0.739**	0.781

Table 2: Comparing four models in detecting positive emotion in customer messages.

For lack of space we report here only the results of a telecommunications company. The full data include 677,936 full interactions (conducted between October and December 2016), with 10,035,328 individual messages. Full interactions include between two and several hundred messages; messages can be customer generated, employee generated, or automatically generated by the service platform (e.g., *"Thank you for your patience. One of our agents will be with you shortly"*). We analyze here only customer messages (mean number of customer messages in an interaction is 12.75, *SD* = 13.33).

For testing Hypothesis 2, we added service quality data collected separately by the company. This included customer assessments of problem resolution and self reported satisfaction with the service. Approximately 50% of the customers were sent a post-service survey (73% of all customers) responded, an acceptable response rate in customer surveys. **Problem resolution** was assessed with a measure known in the service industry as FCR, based on responses to the question "Was your service need resolved in this interaction?" (Yes/No). **Customer Satisfaction (CSAT)** was assessed with the question: "Please rate your satisfaction with the service you received" (responses rated 1-Very unsatisfied to 5-Very Satisfied).

4 FINDINGS

4.1 Customer Emotion Dynamics within Service Interactions

To examine and compare emotion dynamics in different interactions, we standardized length of interactions to 10 sections (or 10 deciles). We then averaged the sentiment of all customer messages in each section, obtaining 10 scores that depict the customer emotion in the section. We used these 10 emotion scores to depict the evolution of emotion in the interaction, and to compare emotions expressed in the first sections to emotions expressed at the end of the interaction. We conducted this comparison for the full dataset, and for a subset of 390, 438 interactions that include 10 or more customer messages. For lack of space we report only the latter[5].

Hypothesis 1, which predicted that service interactions begin with negative emotion and end with positive emotion, was supported. Figure 2 presents the sentiment flow in sections of interactions. In support of Hypothesis 1, a paired-samples t-test confirmed customer emotions at the beginning (the first section) and the end (last section) of interaction as significantly different ($M_{difference} = 0.63$, $t(390437) = 450.52$, $p < 0.001$). For robustness we also compared the first and last two sections, and obtained similar results. The prediction that customer emotions are negative at the start and positive at the end of the interaction, is supported by a single-sample t-test: across multiple interactions, there is more negative customer emotions at the beginning ($t(390437) = -138.72$, $p < 0.001$), and more positive emotions at the end of the interaction ($t(390437) = 458.5$, $p < 0.001$).

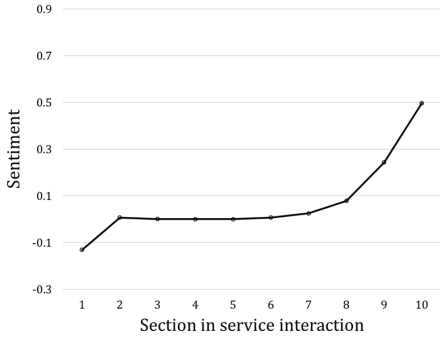

Figure 2: Sentiment flow of service interactions, divided into 10 sections [$n = 390, 438$]

To support the claim that emotions in the beginning of a chat convey service failure while emotions at the end reflect problem resolution, we examined the CustSent engine rules activated in each part. We find the following rules (terms) prevalent in early sections: *error, problem(s), issue(s), wrong, lost, confused, missing, unable, invalid, trouble, cancel, mistake, incorrect.* These terms clearly indicate service failure. For example: *"Something is wrong with my account," or "I have a problem receiving calls."* In contrast, terms that appear more towards the end of service include: *thank(s), good, help, great, works/working, fine, correct, appreciate, nice, happy, best.* These terms more likely indicate service resolution. For example, *"I really appreciate your help," or "That sounds fine. Thanks."*.

4.2 Relating Customer Emotion to Service Quality

We report the following analyses for the subset of 286,671 interactions that include 10 or more customer messages, and whose customers responded to a post-service survey. Hypothesis 2 predicted that the change in customer emotion during a service interaction from negative to positive reflects service quality. To test this hypothesis, we will run a logistic (ordinal) regression to predict FCR (CSAT) from the sentiment score in each section of the interaction.

Resolution of Customer Needs Hypothesis 2 predicted that the evolution of emotion for customers whose issue was resolved is different from the evolution of emotion for customers whose issue was not resolved. To test this hypothesis, we used the section number (a within-subject factor) and the FCR response (as a between-subject factor) in a mixed-effects model as predictors of customer emotion. The interaction between the two factors in this model

[5]To conduct this analysis on all interactions in the data, including shorter interactions, as a robustness check, we stretched interactions with less than 10 customer messages, by duplicating missing quantiles. For example, for an interaction with length 5: 1,2,3,4,5, the 10 points were 1,1,2,2,3,3,4,4,5,5. The results of this "stretched" analysis were similar, and support the robustness of our test.

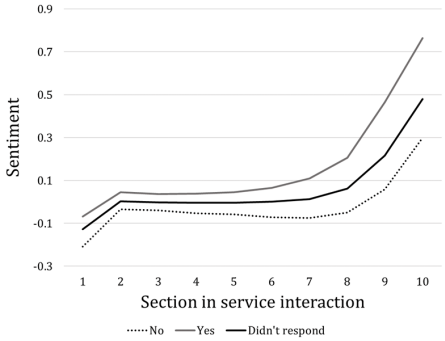

Figure 3: Sentiment in different sections by FCR response [$n = 286,671$]

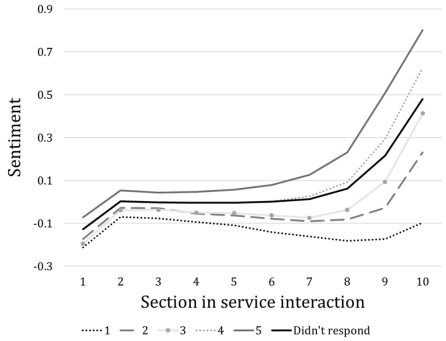

Figure 4: Sentiment in different sections by customer satisfaction (CSAT) response [$n = 286,671$]

indicates whether the evolution in emotion scores differs for different FCR customer groups. We found a significant interaction between section number and customer FCR ($F(9, 1271502) = 2229.12$, $p < 0.001$), which supports that emotions develop differently for customers with different FCR values[6]. Customers who say their issue was resolved have a steeper climb from initial negative emotion, and end with higher levels of positive emotions. Customers who report their issue was not resolved had significantly lower emotion scores at the end of the service (Figure 3). In addition, we support these results with a logistic regression that predicts customer FCR values by customer sentiment scores in each section of the interaction ($\chi^2(10) = 28481.386, p < 0.001$). The model explained 26.1% (Nagelkerke R^2) of the variance in FCR and correctly classified 77.4% of cases. The effect of the sentiment scores in the latter sections was significantly higher than the effect of the early sentiment scores (Beta=.83 and .97 of sections 9–10 vs. Beta=.27 and .07 of sections 1 and 2), which further demonstrate the dependency between emotion dynamics and service outcomes.

Customer Satisfaction Hypothesis 2 also predicted that the evolution of emotion in a service interaction differs between satisfied vs. unsatisfied customers. We test this prediction in a similar analysis, using customer CSAT response as the between-subject factor[7]. A significant interaction between satisfaction and section number ($F(9, 891990) = 3386.85$, $p < 0.001$) again confirmed that emotions evolve differently for customers who end up reporting different levels of satisfaction. Interactions where customers reported a higher satisfaction score had a significantly steeper change in customer emotion during the interaction; the change from the initial negative emotion to the positive at the end was significantly larger (Figure 4).

Here as well, an ordinal regression supported the results, showing that customer sentiment scores in each section of the interaction predicts customer satisfaction scores ($\chi^2(10) = 44725.318, p < .001$). This model explained 28.9% (Nagelkerke R^2) of the variance in customer satisfaction. Importantly, the effect of the sentiment scores in the *latter* sections of customers who reach higher level of satisfaction are significantly larger than the effects of early sections (Beta=.895 and 1.10 of sections 9 and 10, vs. Beta=-1.71 and -1.27 of sections 1 and 2). This again supports the claimed relationship between emotion evolution and satisfaction.

Figures 3 and 4 illustrate the findings of Hypothesis 2, showing patterns of relationships between customer emotions during service and customer evaluations (FCR, CSAT) after the service interaction. The figures also summarize our key theme, showing that interactions that start with negative emotions (because of a service failure), can evolve into (good) service resolution, evident in more positive emotions at the end of the interaction. Our analyses show less influence of initial emotions (early in the interaction), further supporting our interpretation that initial emotions reflect pre-chat service failure. Both figures suggests the existence of a "tipping point" around the middle of the interaction, from which positive customer emotions begin to emerge. We suggest the problem resolution, that may have started around that stage, is connected to this phenomenon.Identifying the exact events that led to such an emergence of customer emotion, and the exact dynamic around it are beyond the scope of our analyses. In short, automated, real time assessments of customer emotion *during* an interaction may essentially replace (more costly and late) evaluations of service quality. Emotion dynamics during the interaction reflect customer satisfaction.

5 DISCUSSION

We introduce a new approach to studying customer emotions in service interactions, and to assessing service quality (service failure and service recovery) in a service interaction. The approach offers a new model for automatic assessment of customer emotions in the service domain, and our analyses provide evidence of the validity of these assessments in identifying customer emotions, and their utility for identifying resolution of customer needs and customer satisfaction. The model allows real-time assessments of customer emotion in spontaneous and real-life service interactions. This new approach has substantial benefits in providing objective, unobtrusive assessments of customer service, that build directly on customers' actual expressions [31], and in assessing customer emotion in far greater granularity than prevailing methods (of

[6]We also found a significant effect of the section number variable ($F(9, 1271502) = 16409.76$, $p < 0.001$), fully supporting Hypothesis 1.

[7]We conducted two analyses, one defined responses of five (5) and one (1) as satisfied and unsatisfied customers, respectively, a second analysis defined responses 1–3 as unsatisfied, and 4–5 as satisfied. The results were identical.

customer surveys) can depict. Current practices typically aggregate reports of customers into bins of "satisfied" and "not satisfied". Our approach offers a more complete picture, of the changes in customer emotion during an interaction, showing the relation of these changes to service quality evaluations. This approach can be used to detect service delivery issues in real time, allowing interventions as a problem occurs, rather than after it occurs, which is currently the common practice. To this end, the CustSent model is applied in the LivePerson chat service platform, and monitors real-time customer emotion development in many different brands.

5.1 Contributions

Our paper makes three core contributions: (a) methodologically, it proposes automated sentiment analysis as a useful tool for both service delivery and service research ; (b) theoretically, it documents the meaning of trends and changes in emotions that occur within an individual interaction; (c) managerially, it suggests a new way to leverage sentiment analysis to improve service operations. Our approach suggests a wide range of ideas that can promote research and management of service delivery [22, 23], operations [10], and human resource management [17].

Our results show that the level of positive emotions that customers reach (compared to where they start) reflects the quality of the service interaction. Better service quality is evident in the (positive) emotions customers express in the latter part of their interactions. Thus, automated emotion assessment, conducted in real-time during service interactions, can be used to evaluate service quality, and to intervene toward better service recovery [6]. Identifying customers whose emotions do not improve towards the end of the interaction can help managers intervene before a service situation escalates. A system of alerts, for example, when customer sentiment stays negative, can be used as notification that something is wrong. In addition, our prediction model can be used for developing measurements to replace customer surveys, using an automated objective tool, rather than post-hoc subjective assessments.

5.2 Limitations and Future Research

A natural step for future research is assessing when and what emotion alerts should be activated, and what their impact may be. There are also limitations to our work that call for more research. First, our analyses currently detect customer emotion, only; a desirable extension for a complete picture of service interactions would be monitoring expressions of agents. Employees must regulate the emotions they express toward customers, performing what [25] described as "emotional labor". A separate emotion detection tool is therefore needed for analyzing employee emotion. Second, the approach we suggest can help investigations of effects of customer emotion on employee performance. Some recent research, for example, shows that customer sentiment influences employee response time and employee tendency to take unscheduled breaks [2, 4]. Dynamic planning of time allotted to a given service interaction, or of employee breaks based on identified customer emotions, can build on these analyses, and help reduce employee burnout. Third, we analyzed only customer textual expressions toward detecting customer emotions. Sentiment analysis tools may be improved with integration of additional aspects of customer behavior, like key strokes, or engagement history (cf. https://www.clicktale.com/). Such integration can potentially improve predictions of service evaluations. Lastly, combining sentiment analysis with aspect analysis (e.g. [5]), in the context of service delivery, can further distinguish the emotions resulting from service failure and recovery; in addition this may also provide the business some guidance into optimizing service recovery strategies. This opens up numerous opportunities for research.

ACKNOWLEDGEMENT

We thank Naama Tepper and Shlomo Lahav for initiating the collaboration between the Technion and LivePerson, Ella Nadjharov, Igor Gavako and Dr. Valery Trofimov (the SEELab team at the Technion), and the following students for helping CustSent testing and evaluation: Galia Bar, David Spivak, Gabby Mayer, Cassidy Laidlaw, Laura Blumenfeld, Beaux Ballard.

REFERENCES

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media*. Portland, Oregon, 30–38.
[2] Daniel Altman. 2017. *Modeling employee behavioral reactions to emotions expressed by customers.* Master's thesis. Technion—Israel Institute of Technology.
[3] Teresa M Amabile, Elisabeth A. Schatzel, Giovanni B Moneta, and Steven J Kramer. 2004. Leader behaviors and the work environment for creativitity. *The Leadership Quarterly* 15, 1 (2004), 5–32.
[4] Shelly Ashtar. 2017. *The effect of customer emotion and work demands on employee unscheduled breaks.* Master's thesis. Technion—Israel Institute of Technology.
[5] Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 804–812.
[6] Riza Casidy and Hyunju Shin. 2015. The effects of harm directions and service recovery strategies on customer forgiveness and negative word-of-mouth intentions. *Journal of Retailing and Consumer Services* 27 (2015), 103–112.
[7] Marie T. Dasborough. 2003. Cognitive Asymetry in employee affective reactions to leadership behaviors. *Leadership Quarterly* 17, 2 (2003), 1–32.
[8] Tom DeWitt, Doan T. Nguyen, and Roger Marshall. 2008. Exploring customer loyalty following service recovery: The mediating effects of trust and emotions. *Journal of Service Research* 10, 3 (2008), 269–281.
[9] Susan Folkman, Richard S. Lazarus, Christine Dunkel-Schetter, Anita DeLongis, and Rand J. Gruen. 1986. Dynamics of a stressful encounter. *Journal of Personality and Social Psychology* 50, 5 (1986), 992.
[10] Gerard George, Ernst C Osinga, Dovev Lavie, and Brent A Scott. 2016. Big Data and Data Science methods for management research. *Academy of Management Journal* 59, 5 (2016), 1493–1507.
[11] Markus Groth and Alicia A Grandey. 2012. From bad to worse: Negative exchange spirals in employee-customer service interactions. *Organizational Psychology Review* 2, 3 (2012), 208–233.
[12] Jonathan Herzig, Guy Feigenblat, Michal Shmueli-Scheuer, David Konopnicki, and Anat Rafaeli. 2016. Predicting customer satisfaction in customer support conversations in social media using affective features. *Proceedings of UMAP '16* (2016), 115–119.
[13] George Lakoff. 1984. Performative subordinate clauses. In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, Vol. 10. 472–480.
[14] Robin Lakoff. 1976. Language in context. *Language* 48, 4 (1976), 907–927.
[15] Rikard Larsson. 1993. Case survey methodology: Quantitative analysis of patterns across case studies. *Academy of Management* 36, 6 (1993), 1515–1546.
[16] Christopher Manning, Raghavan Prabhakar, and Schütze Hinrich. 2009. Introduction to information retrieval. (2009).
[17] Samuel T McAbee, Ronald S Landis, and Maura I Burke. 2017. Inductive reasoning: The promise of Big Data. *Human Resource Management Review* 27, 2 (2017), 277–290.
[18] Janet R McColl-Kennedy and Amy K Smith. 2006. Customer emotions in service failure and recovery encounters. In *Research on emotion in organizations*, W.J. Zerbe, N.M. Ashkanasy, and E.E.J. Haertel (Eds.). Vol. 2. Emerald Group Publishing Ltd, Bingley, UK, Chapter 10, 237–268.
[19] Janet R McColl-Kennedy, Beverley A Sparks, By Beverley Sparks, and Janet McColl-Kennedy. 2003. Application of Fairness Theory to Service Failures and Service Recovery. *Journal of Service Research* 5, 3 (2003), 251–266.

[20] Ella Miron-Spektor, Dorit Efrat-Treister, Anat Rafaeli, and Orit Schwarz-Cohen. 2011. Others' anger makes people work harder not smarter. *Journal of Applied Psychology* 96, 5 (2011), 1065–1075.

[21] Lakshmi Narayanan, Shanker Menon, and Paul E Spector. 1999. Stress in the workplace: A comparison of gender and occupations. *Journal of Organizational Behavior* 20, 1 (1999), 63.

[22] Francisco Villarroel Ordenes, Stephan Ludwig, Ko De Ruyter, Dhruv Grewal, and Martin Wetzels. 2017. Unveiling what is written in the stars. *Journal of Consumer Research* 43, 6 (2017), 875–894.

[23] Anat Rafaeli, Daniel Altman, Dwayne D Gremler, Ming-Hui Huang, Dhruv Grewal, Bala Iyer, A. Parasuraman, and Ko de Ruyter. 2017. The future of frontline research: Invited Commentaries. *Journal of Service Research* 20, 1 (2017), 91–99.

[24] Anat Rafaeli, Amir Erez, Shy Ravid, Rellie Derfler-Rozin, Dorit Efrat Treister, and Ravit Scheyer. 2012. When customers exhibit verbal aggression, employees pay cognitive costs. *Journal of Applied Psychology* 97, 5 (2012), 931–950.

[25] Anat Rafaeli and Robert I Sutton. 1987. Expression of emotion as part of the work role. *Academy of Management Review* 12, 1 (1987), 23–37.

[26] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 1631–1642.

[27] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics* 37, 2 (2011), 267–307.

[28] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29, 1 (2010), 24–54.

[29] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, and Di Cai. 2010. Sentiment strength detection in short informal text. *The American Society for Informational science and technology* 61, 12 (12 2010), 2544–2558.

[30] Gerben A van Kleef, Carsten K W De Dreu, and Antony S R Manstead. 2004. The interpersonal effects of anger and happiness in negotiations. *Journal of Personality and Social Psychology* 86, 1 (2004), 57–76.

[31] Eugene J Webb, Donald T Campbell, Richard D Schwartz, and Lee Sechrest. 1966. *Unobstrusive measures: Nonreactive research in the social sciences*. Vol. 111. Rand Mc Nally, Chicago.

[32] Howard M Weiss and Russell Cropanzano. 1996. Affective Events Theory. *Research in Organizational Behavior* 18, 1 (1996), 1–74.

[33] Renxian Zhang, Dehong Gao, and Wenjie Li. 2011. What are tweeters doing: Recognizing speech acts in Twitter. *Analyzing Microtext* (2011), 86–91.

[34] Shaodian Zhang, Erin Bantum, Jason Owen, and Noémie Elhadad. 2014. Does sustained participation in an online health community affect sentiment?. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association.