Emotion in Text-Based Customer Service: Using Automatic Emotion Detection to Identify Trends and Relationships

Galit B. Yom-Tov, Anat Rafaeli, Shelly Ashtar
gality@technion.ac.il, anatr@ie.technion.ac.il, shellya@campus.technion.ac.il
Technion—Israel Institute of Technology

Daniel Altman, Monika Westphal
{altmand, westphal}@campus.technion.ac.il
Technion—Israel Institute of Technology

Michael Natapov, Neta Barkay
{michaelna, netab}@liveperson.com
LivePerson Inc.

Recent developments in Machine Learning and Natural Language Processing enable the development of automated Sentiment Analysis tools. Currently available tools were developed to identify emotions in well structured texts such as product or film reviews. We extend this approach to develop a new tool for objective assessment of customer emotions in spontaneous interactions between service agents and customers. We then use this tool, to explore customer sentiment using data of real customer-agent interactions in three companies from different industries. Our analyses shed light on the presence of positive and negative emotion in customer service, and identify patterns of customer emotion over the course of service interactions. The analyses also quantify the link between customer emotions and three accepted indices of customer service, and document different patterns of emotion between effective and less effective interactions. Our tool and analyses provide the basis for incorporating sentiment measurements into operational decisions. Throughout the paper we discuss implications of using automated sentiment analysis tools for the management of service centers.

1. Introduction

With the growing role of service in the modern economy, understanding the effects and dynamics of emotions expressed during service, is critical. Service interactions involve an exchange of information, critical operational dilemmas and economic outcomes, as well as inter-personal social dynamics (McColl-Kennedy and Smith 2006). Technology allows modern-day companies to replace

traditional service encounters (face-to-face, telephone) with technology-mediated service encounters (Massad et al. 2006, van Dolen and de Ruyter 2002), which allow customers and service employees to be in different physical locations, and connect via a technological interface (Schumann et al. 2012, Froehle and Roth 2004). Voice services are well researched (e.g., Gans et al. 2003). Services can also be delivered using written messages, for example when employees and customers chat through corporate websites, Twitter or Facebook. The current paper is on chat as a service platform (cf., `http://LivePerson.com`).

Regardless of the medium of customer service interactions, service interactions are social interactions, and inevitably comprise emotional dynamics. Organizations strive to satisfy customers (cf., Oliver et al. 1997), and to avoid customer anger (Schneider and Bowen 1999, McColl-Kennedy et al. 2009). Service employees are instructed to display certain emotions as part of their job (cf., Rafaeli and Sutton 1987), because of the assumption that emotion displays improve customer satisfaction and increase sales. Yet available research on emotions in service has been constrained in multiple ways. The first key contribution of the current paper is overcoming these constraints.

A first category of constraints that we overcome regards tools for measuring emotions. We introduce the use of automated assessments of customer emotion, which is objective and allows analyses of larger data sets. Previous research was constrained by available tools, and researchers had to rely on self-report measures (Donaldson and Grant-Vallone 2002, Paulhus and Vazire 2007), or on observations (Pugh 2001, Rafaeli and Sutton 1990, Sutton and Rafaeli 1988). Self-reports (i.e., employee or customer responses to written surveys about how they feel) and observations (i.e., recordings of people's feelings as documented by research assistants) are known to provide a picture of limited accuracy, with multiple biases (cf., Donaldson and Grant-Vallone 2002, Howard and Dailey 1979). Our analysis overcomes these limitations by developing objective assessments of emotion. A second constraint regards the research design, where research on emotion in service has relied primarily on scenarios and laboratory studies; the external validity of available findings has therefore been limited. These methods are useful for providing a basic understanding of emotional dynamics. An essential extension, and a major contribution we make, is testing dynamics of customer emotion in actual and live customer service interactions.

A second contribution of the present paper is in the breadth of data analyzed. The constraints of available tools and research paradigms for the study of service interactions have meant that available studies and findings are based on small samples in a limited service context. For example, Pugh (2001) reports on customer emotions based on data of 220 bank customers who agreed to participate in the study sampled in a single week. Similarly, Wirtz and Mattila (2003) report on customer feelings based on responses of what they identified as "a convenience sample of 187 working adults" to surveys. Smith and Bolton (1998) report on customer satisfaction based on responses of 275 undergraduate students in one study, and 520 customers of one hotel in a second study. Joireman et al. (2013) report on customer reactions based on surveys of 250 customers. All these studies are very important in suggesting novel effects of emotion. However, results based on such small samples may embed various extraneous effects, and it is not clear to what extent they reflect actual emotional dynamics in large scale service delivery. The present paper contributes by analyzing large samples of real-life customer service interactions.

A third contribution we offer is the measurement of emotion as it occurs throughout service encounters. Previous research rarely reported results about dynamics that occur during service encounters. Available theory clearly indicates that service interactions in organizations envelop multiple manifestations of emotions (cf., Affective Events Theory; Weiss and Cropanzano 1996, Elfenbein 2007). Some research shows that emotional incidents are important to recognize because they influence various factors (Weiss and Beal 2005, Wegge et al. 2006). The unique data that we analyze, and the tools we use, provide insights into the flow of emotions throughout service interactions. Moreover, our analyses will relate the flow of emotions during interactions to the outcomes of the interactions.

An important progress that allows us to make these three contributions is the development of a tool for automated sentiment analysis, that build on developments in computational linguistics (cf., Pang and Lee 2008, Taboada et al. 2011, Tausczik and Pennebaker 2010, Thelwall 2013). We suggest here that models for automated emotion detection allow empirical investigation of

emotions in spontaneous, real-life customer service interactions. Service platforms increasingly include archives of customer expressions and employee behaviors. We suggest that combining these rich data archives with sentiment analysis tools provides a new approach to assessing emotion in service delivery. This methodology also allows connecting emotions to key operational aspects such as employee effectiveness and customer satisfaction. Linking automatic emotion detection to various indices of service performance, affords multiple managerial insights, and can help address important questions about the effects of customer emotions on service interactions (Rafaeli et al. 2016, Yom-Tov et al. 2017a).

We begin by describing a new tool that we developed for automatic detection of emotions in customer messages in chat service interactions (§2). The tool (which we call CustSent, short for Customer Sentiment detection), relies on a lexicon-based model, that combines word classification with an added layer of Natural Language Processing (NLP). We had to develop a new model because text of natural and spontaneous service interactions is challenging to available sentiment analysis tools. This text includes natural, unedited language, and comprises several lines, and often short lines, including many lines comprising only a few words, such as "sure", or "no, thanks". Real-life interactions are dynamic and vary in length and context, and are very different from the type of text used in the development of known sentiment analysis models (cf., Buechel and Hahn 2017, Strapparava and Mihalcea 2007, Thelwall et al. 2010).

We evaluate the performance of CustSent by comparing its identification of emotion in customer interactions, to other prevalent sentiment analysis solutions, including LIWC (Tausczik and Pennebaker 2010), Stanford (Socher et al. 2013) and SentiStrength (Thelwall et al. 2010). Our comparisons show that LIWC and Stanford do not perform well in chat services, while SentiStrength and CustSent are comparable in identifying positive customer emotion, but CustSent is significantly superior in recognizing negative customer emotions. Negative emotions are critical to the context of customer service, as indicators of service failure and customer dissatisfaction. Our report here presents CustSent as a more effective tool for studying customer emotions in service operations

than other, available sentiment analysis tools. Section 3 reports on the data we are analyzing, and Section 4 reports findings about emotions occurrence and dynamics in typical service interactions. We conclude in Section 5, with a discussion of the implications of our idea of automated sentiment analysis of customer service interactions for continued service research.

In short, the paper provides first a tool for monitoring emotions within text-based service interactions, and second, a unique glance into what these emotions look like in reality. For example, we observe that in contrast to common belief, clear negative emotions are expressed by only a small proportion of customers ($< 10\%$), and appear in less that 5% of customer sentences. Positive emotions are much more common, and are expressed in higher intensity (§4.1). Both positive and negative emotions appear evenly over employee's work shifts (§4.4). While investigating the dynamic of emotions during service interactions, we identify an empirical structure of the emotional roller-coaster inherent to service interactions between employees and customers. We observe that the emotional footprint of a typical interaction has an S shape: Initial emotions are negative, which segue into a relatively long segment of little or no emotions, and an increase towards positive emotion in the end of the interaction (§4.2). We quantify for the first time the connection between such customer emotional dynamics and common measures of customer satisfaction (§4.3). We observe that the beginning of interactions do not differ between satisfied and unsatisfied customers, while significant differences emerge after a "tipping point" that occurs somewhere during the interaction, and lead to an improvement in customer emotion from the middle to the end of the interaction. This has implications for the use of sentiment engines, such the one developed here, both for monitoring service success in real time and as an objective alternative to biased satisfaction surveys. We discuss such implications throughout the manuscript.

2. Methods

2.1. Automatic assessment of customer emotion in text-based service interactions

Automated sentiment analysis is managerially useful because it can allow assessment of emotion in large-scale customer data, and can pave the way to real-time information about customer emotion. Managers of service systems can greatly benefit from such information, because it provides a clear

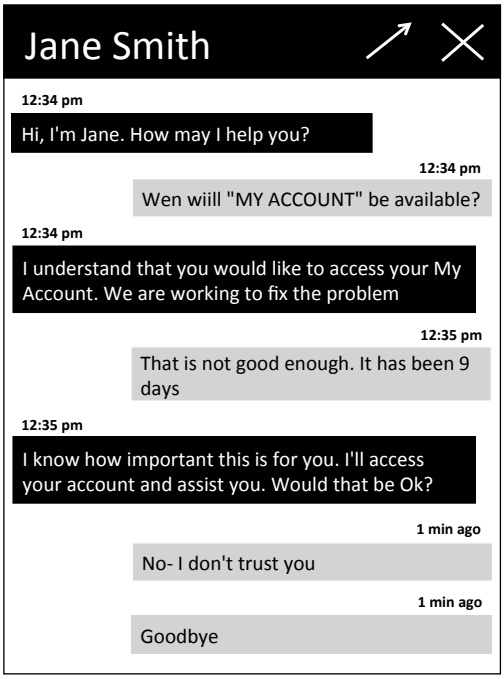
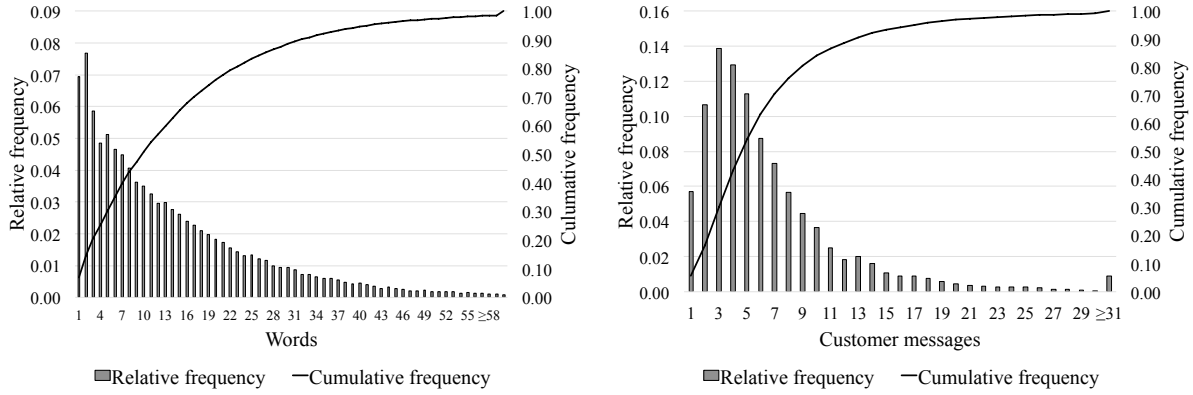


Figure 1 An example of service interaction between employee and customer

indication of how their customers feel about the service. Our current report regards customer emotion in text-based (chat) service interactions. These service interactions comprise a sequence of interdependent messages between customers and service employees (See Figure 1). A message becomes visible to the second party (and is generated as an entity in the system) when the message author (the employee or the customer) hits the "Enter" button. A message may include anywhere from a single word or symbol to hundreds of words, as well as punctuation marks (e.g., commas and periods), and emoticons. Complete service interactions may comprise as few as two messages and as many as hundreds of messages, generated by customers, service employees and automated system responses. Figure 2 presents distributions of the number of words in customer messages, and of the number of customer messages in 25,714 full service interactions of a large North American airline transportation company conducted over 3 months.

Service interactions typically consist of multiple customer messages, so they can be described at the atomic level of individual messages (i.e., identifying emotion in each individual customer message), and at a cumulative level of full interactions (i.e., identifying emotion in a full interaction). Analyses at the two levels serve different functions. Identifying emotion at the message level



(a) Number of words in customer messages

(b) Number of customer messages in service interactions

Figure 2 Distributions of number of words in customer messages, and number of customer messages in service interactions [Transportation, $n = 25,714$]

enables real-time detection of a customers emotional state; monitoring the cumulative emotion at the full interaction level is useful for assessing overall customer satisfaction or as an indicator of the performance of service employees. Both types of analyses require a tool that can perform efficient and accurate detection of customer emotion.

2.1.1. The need to design a new model for detecting emotion in service interactions

To identify the emotions in service chats, we began by testing the ability of available state-of-the-art sentiment analysis tools to detect emotion in customer messages. Among others we tested the Stanford RNTN (Socher et al. 2013), SentiStrength (Thelwall et al. 2010), and LIWC (Tausczik and Pennebaker 2010). As we describe next, these standard tools could not provide assessments that are satisfactorily valid. Therefore we developed and validated a special tool (*CustSent*), which automatically detects emotions in customer messages (within an interaction). Below, we describe the way the tool was designed and how it operates.

Our assessment of the sentiment detection relied on two standard metrics used in evaluating accuracy of sentiment identification: ***precision*** and ***recall***; these metrics evaluate predictions of whether an element belongs to a subclass of a population (Powers 2011). ***Precision*** is the proportion of retrieved instances that are relevant, and ***recall*** is the proportion of relevant instances

that are retrieved (Powers 2011, Manning et al. 2008). In our context, to measure precision and recall of the negative emotion class, one compares the number of messages detected as negative by a sentiment analysis tool to the number of messages coded as negative by human judges. Formally, denote α_{neg} as the number of messages detected as negative by a sentiment analysis tool, β_{neg} as the number of messages coded as negative by human judges, and γ_{neg} as the number of messages detected as negative by a sentiment analysis tool and coded as negative by human judges. Therefore:

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

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

Precision and recall of the class of positive emotion are similarly defined. Precision measures the trustworthiness of an identified emotion, and recall measures the extent to which all emotions present in a text are recognized. In analyses of customer service, most important is the accurate identification of negative emotion because this can indicate a service failure, and the quality of the service provided by service employees (cf., Joireman et al. 2013, Groth and Grandey 2012). Both uses require that false alarms (instances where an alert of a customer negative emotion is lit and a service failure did not occur) should be kept to a minimum.

Our preliminary tests showed that available tools for identifying negative emotions have low precision. We believe this is because the text of service interactions is unique and significantly different from standard corpora that have been used for sentiment model training. Specifically, movie reviews are a common resource, and were used for developing the Stanford Sentiment Treebank (Socher et al. 2013). Reviews typically include unambiguous, straightforward opinions, and a comprehensive description of issues. In contrast, service interactions typically comprise short sentences, do not necessarily maintain a coherent text 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, all of which may relate to the emotion of the writer, and are likely to be missed or misinterpreted by available sentiment detection models (Boiy et al. 2007).

Consequently, a special emotion detection model is needed for the accurate identification of customer sentiment in short, naturally expressed customer messages in the context of customer service interactions. There may be variations in the nature of customer service interactions of different industries, or in the customer interactions of different organizations within the same industry, depending on the service culture of the organization (Schneider 1990). Our goal was to create a robust model, that would afford optimal performance across multiple interactions and different industries. Following (Taboada et al. 2011), we decided to base the design of the model on the lexicon-based and rule-based approach; this allows us to leverage analyses of a large amount of archival interaction data, and develop easier adaptation to different domains. The alternative of a machine learning approach for developing the model would have required training a separate model for each service domain, and would have implied a very high annotation cost.

2.1.2. CustSent: The new emotion detection model The model we developed (CustSent) is for chat interactions conducted in English (UK and USA English), and assigns a score to each customer message by applying a set of rules. Each rule assigns a numeric integer score to words or nonverbal elements of the message; the score may be zero, positive or negative. A score of zero indicates no emotion, positive scores indicate positive emotion, and negative scores indicate negative emotion. The magnitude of the score indicates the emotion intensity. Scores of multiple rules are then aggregated, creating an overall emotion score for each message. Message emotion scores are theoretically unbounded, and practically, 99% of customer-message scores are between -3 and +3 (the distribution of the scores is presented in §4.1). Total message scores above zero indicate positive sentiment, and scores strictly below zero indicate negative sentiment. A value of zero indicates no emotion or an equal amount of positive and negative emotion in the message.

Two types of rules determine the emotion score. One set of rules, described below as "Lexicon Based Rules", assigns a ***base score*** to emotionally charged words (***anchors***); these are manually annotated words that comprise lexicons of different base polarity and intensity; e.g., positive words: *excellent, great, works*, and negative words: *horrible, confused, cancel*. These rules include

adjustments for the ***context of the anchor***, which is defined as the presence of negation or/and intensification words in three words preceding an anchor[1]. An anchor that appears without negation and/or an intensifier is assumed to be ***without a context*** and its score remains unaltered (remains the base score of the anchor). The context-defined scores of anchors in a message are added up to create the preliminary score of a message.

A second set of rules, described below as "Sentence Level Rules" updates the preliminary score of the message, based on non-verbal features, such as exclamation or question marks and emoticons, and verbal features such as polite words (e.g. *sorry*), appreciation words (e.g. *thanks*) or popular abbreviated slang (e.g. *LOL* or *lol*—an acronym for *laughing out loud*). Both the lexicons and the sentence level rules were derived inductively by looking through large samples of customer interaction data. We next describe these rules explicitly, and provide examples for clarity.

2.1.3. Lexicon rules: Assigning emotion scores to anchors Our model uses five lexicons with different levels of the base sentiment polarity: ***negative*** (base score -1), ***very negative*** (-2), ***positive*** (+1), ***very positive*** (+2) and ***weak positive*** (base score 0, but becomes negative if negated). Each lexicon has its own ***context score shift***—a pattern of polarity or intensity shifts of the base score under negation and intensification.

The first four lexicons—negative, very negative, positive, very positive—follow similar patterns of context score shift:

Intensification words intensify the base score of an anchor by 1 point

$$\textit{pleased}\ (+1) \rightarrow \textit{very pleased}\ (+2)$$

$$\textit{excellent}\ (+2) \rightarrow \textit{absolutely excellent}\ (+3)$$

$$\textit{disappointed}\ (\text{-}2) \rightarrow \textit{extremely disappointed}\ (\text{-}3)$$

Negation words shift the base polarity of an anchor by 2 points in the direction of the opposite polarity

[1] We compared a model with 2, 3, 4 and 5 preceding words and found 3 words to be optimal in English interactions. This process is language dependent. For example, in French we find a different number of words to be relevant, and find words after the anchor to be critical for the analysis.

pleased (+1) → *not pleased* (-1)

excellent (+2) → *not excellent* (0)

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 ***weak positive*** lexicon is different from the above four lexicons, comprising words that are neutral without a context and become negative with a negation:

enough (0) → *not enough* (-1)

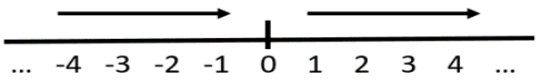
like (0) → *don't like* (-1) → *really don't like* (-2)

The terms *like* and *don't like* exemplify the importance of developing a model tailored for customer service interactions. The word *like* is considered positive in available lexicons, yet a positive rating of this word in our texts would be erroneous. In developing the engine we examined simple frequencies of all words, and found less than 10% of the appearances of the word *like* without a context, to be positive. The most common usage of the no context *like* is in a neutral construct of *"I would like to "*. However, when the word *like* is negated, it almost always has a negative connotation, as in *"I dont like "*. This complexity places the term *like* in the weak positive lexicon of our model.

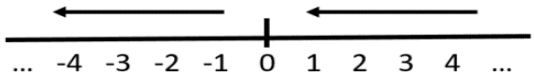
2.1.4. Sentence level rules: Updating emotion scores from anchors to scores for full messages The context-defined scores of anchors in a message are added up to create the preliminary score of a message. For accurate indication of overall emotion in a message, additional analyses of complete sentences are essential. Sentence level rules assess features such as sentence punctuation, emoticons, special language, and special structure. Sentence level rules are not mutually exclusive. So multiple rules can be implemented for any given sentence.

As summarized in Table 1, sentence level rules can augment a sentence score in four ways:

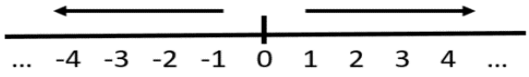
Adding: shifting a score in a positive direction



Subtracting: shifting a score in a negative direction



Strengthening the intensity: increasing the extent to which a score is different from zero



Weakening the intensity: decreasing the extent to which a score is different from zero

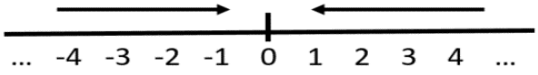


Table 1 **Sentence level rules for modifying the emotion score created by lexical words.**

Rule name	Update a message score by
Question	Weakening the intensity by 1
Politeness	Weakening the intensity by 1
Condition	Weakening the intensity by 1
Positive slang	Adding a factor of 2
Smiley	Adding a factor of 1
Frowny	Subtracting a factor of 1
Negative idiom	Subtracting a factor of 2
Thanking phrase	Adding a factor of 1, 2, or 3 depending on specific phrase (thx, thank you, thank you very much, etc.)
Multiple punctuation with non-zero score	Strengthening the intensity by 1
Multiple punctuation with zero score	Subtracting a factor of 2

Question rule: A question structure has a different emotional load than a declarative sentence with the same wording (Lakoff 1972, 1984, Zhang et al. 2011), because questions reduce the intensity of any emotion expressed. For example, consider the following two sentences, scored as negative and neutral, respectively:

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)

The identification of questions requires special attention, since customers do not always bother typing question marks ("?"). The presence of any word from a list of question words (e.g. *why, where, does, is* etc.) at the beginning of a sentence, or a question mark at the end of a sentence, weaken the intensity of the sentence sentiment score.

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, we subtract from the intensity of a sentence sentiment score in presence of politeness and/or conditioning.

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 rule therefore adds to the sentence sentiment score in presence of such slang words (e.g., *lol, lmao, haha*).

Smilies and Frownies: Smilies, e.g., *:-)* and frownies, e.g., *:(* are clear, non verbal indicators of emotions. They add to or subtract from the sentence score, respectively.

Negative idioms: Some stable phrases/idioms implicitly convey emotion because of the associations they insinuate. Some examples include *"been waiting"*, *"fed up"*, or *"your fault"*. These and similar idioms subtract from the sentence sentiment score:

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

Thank-you phrases: Phrases conveying thanks of the customer add a positive factor to the sentiment score of a message in which they appear. The positive factor depends on the extent to which extreme thanks are conveyed:

no, thanks (+1)

thanks a lot! (+2)

thank you sooo much for your help! (+3)

Multiple punctuation: Another common feature that demands special treatment is multiple exclamation and/or question marks. Inductive analyses led us to model several patterns. Thus, a non-zero sentiment score is intensified by multiple punctuation marks:

I am *confused* (-1) → I am *confused!!!* (-2)

A positive (e.g., thanks phrase) score is similarly intensified:

thanks (+1) → *thanks!!!* (+2)

And a neutral sentiment becomes negative:

hello (0) → *hello????* (-2)

2.2. Assessing the accuracy of the CustSent model

A common practice for validating sentiment analysis models is to use a predefined Gold Standard corpus (see e.g., Nakov et al. 2016, Pang and Lee 2004, §3). However, as noted, customer service interactions are different linguistically and grammatically from typical texts, and there is no comprehensive gold standard corpus for customer service interactions. Generating it would demand a significant annotation effort. As an alternative validation strategy we decided to estimate the accuracy metrics from samples of customer messages (see Bommannavar et al. 2014, for a discussion on precision and recall estimation from samples). We created a sample of customer messages and asked human judges to code the emotions in it. In this section, we first describe the coding process and then explain how the sample for coding was built. We then compare this coding to the emotions detected by CustSent, thereby estimating the accuracy metrics of CustSent.

2.2.1. Coding emotions in customer messages by human judges We asked a group of native English speakers to read through a random sample of service chats, and together (the researchers and this group) developed a protocol for coding emotions in such messages. We then recruited a second group of people and asked them to use this protocol to code emotions in a different sample of customer messages. Both groups were unfamiliar with the model rules.

In the first stage, 6 native English speakers read through a random sample of 50 interactions taken from four companies in distinct industries (telecommunication, retail, entertainment and technical support). They identified positive and negative emotions in the 1,439 customer messages comprising the 50 interactions. The coders worked as two teams of three people identifying positive, and three identifying negative emotion, in an iterative process that took approximately 35 hours.

The coding was done on each message, incrementally, based on reading the entire conversation until that message. The goal of this stage was to develop a protocol for coding emotions in customer service messages. The protocol is available in the E-companion file.

In the second stage, the protocol was given to a second group of native English speakers, who were asked to use it to code a different sample, as described below. Two new teams of three coders each coded each message, one coding positive, the other coding negative emotion. Each team was assigned a psychology graduate student who had expertise in the topic of emotions in customer service, and helped resolve disagreements. The coding scores were determined by consensus resolution (Dasborough 2006, Amabile et al. 2004, Larsson 1993, Narayanan et al. 1999), with differences in potential coding discussed before all three coders in the group agreed on the final assignment of scores. Thus, the final agreement of the coding was in effect 100%.

2.2.2. The coding sample The data coded in this second stage required a unique form of sampling of customer messages. This is because the initial steps of coding revealed that sampling random messages leads to a majority (∼70%) of messages containing no emotion. Therefore, a stratified approach of sampling customer messages was used, in which the extracted sample of messages is biased to include a lower proportion of non-emotional messages than a random sample. Specifically, we split all messages to the 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. This sampling procedure allows us a straightforward estimation of CustSents precision in the negative and positive classes, following Formula (1). To evaluate CustSents recall and the performance metrics of other engines, we adjust Formula (1) and Formula (2) by assigning a weight to each message. The weight of each sampled message 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. It follows that a message from the i-th stratum has the weight $w_i = N_i/(N_1 + N_2 + N_3)$. Then, each message coded as negative by the human judges contributes its weight to the precision and recall formulas. The same is true for each message detected by model M (including CustSent) as positive.

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 of identifying negative emotion by the model M is now:

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

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). Formula (2) for recall of a model M on negative class is modified as follows:

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

Precision and recall on the positive class are adjusted in a similar fashion.

For the final assessment of the validity of CustSent we followed these procedures to sample customer messages from service conversations conducted during the first week in March 2016, with two firms, a 'Telecommunication' (see Table 2) and a 'Retail' firm (see Table 3). We aimed for a sample of 300 messages from each company, with 100 positive messages, 100 negative messages, and 100 no-emotion messages as assessed by CustSent. Due to technical issues, the final effective sample comprised 597 customer messages that were coded by the human coders. Tables 2 and 3 summarize the sample, the strata they represent and the corresponding weights.

Table 2 Data used for assessing the accuracy of the CustSent engine—Customer Interactions in a Telecommunication Service Organization.

Stratum—sentiment polarity detected by CustSent	Stratum size	Sample size	Weight of a sampled message
Negative	54,671	100	0.087
Neutral	533,958	100	0.854
Positive	36,977	99	0.059

Table 3 Data used for assessing the accuracy of the CustSent engine—Customer Interactions in a Retail Service Organization.

Stratum—sentiment polarity detected by CustSent	Stratum size	Sample size	Weight of a sampled message
Negative	20,003	99	0.056
Neutral	303,531	99	0.852
Positive	32,828	100	0.092

2.2.3. Accuracy of CustSent compared to other models To assess the accuracy of CustSent, we report its precision, recall and F-measures. F_1 is a standard way to aggregate precision and recall into one metric:

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

F_1 is a harmonic average of precision and recall, and assigns similar weight to both metrics. Since in the customer service context our main goal is precision, we also deploy the $F_{0.5}$ metric, a variation of F_1, in which precision is weighed twice as important as recall (Manning et al. 2008):

$$F_{0.5} = \frac{(1+0.5^2) \times Precision \times Recall}{0.5^2 \times Precision + Recall}$$

All these metrics – Precision, Recall, F_1, $F_{0.5}$ – are calculated separately for the negative and positive emotion classes, and for the combined emotion class, which measures the ability to detect any emotion in a message. We also calculate the corresponding metrics for the Stanford Sentiment Analysis RNTN model (Socher et al. 2013), the LIWC tool that is commonly used in social science research (Tausczik and Pennebaker 2010), and SentiStrength (Thelwall et al. 2010)[2].

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 4; $p < 0.0001$[3]). Results concerning the recall are less clear: Two engines (LIWC and SentiStrength)

[2] We first calculated the metrics separately for each of the firms. The results were not substantially different. This suggests the robustness of the CustSent engine, for detecting customer emotions in different service contexts and industries. For the sake of brevity we present only the metrics on the combined sample of 597 messages, with each message weighted corresponding to its proportion in the population.

[3] All reported p-values refer to a comparison of CustSent to the best result in the same category.

have lower recall than CustSent. The Stanford engine has a higher recall of negative emotions, but its precision is extremely low; the Stanford engine cannot be considered a viable option for the purpose of identifying emotions in the context of customer service since it does not identify $\frac{2}{3}$ of the cases of negative customer emotions. The $F_{0.5}$ value of CustSent is the highest among the compared detection models.

Table 4 **Comparing four models in detecting negative emotion in customer messages.**

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

In the assessments of positive emotions, the precision of CustSent is generally superior to 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 the SentiStrength model (Table 5).

Table 5 **Comparing four models in detecting positive emotion in customer messages.**

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 6 summarizes the ability of all models to detect ***any*** emotion (either positive or negative) in a message. In addition to precision, recall and F-measures, we also report ***accuracy*** (Manning et al. 2008), which is the proportion of correctly detected presence or lack of emotion in the sampled messages. Here again we see that CustSent is significantly better than the other engines ($p < 0.02$) in precision, and comparable with Sentistrength in all other metrics.

In summary, our analyses confirm that the CustSent engine we developed satisfies common criteria for assessing negative, positive and overall emotion in written customer communication.

Table 6 Comparing four models for detecting any emotion in customer messages.

Model	Precision	Recall	F_1	$F_{0.5}$	Accuracy
CustSent	0.832	0.45	0.584	0.711	0.721
Stanford	0.395	0.42	0.407	0.399	0.526
LIWC	0.521	0.481	0.5	0.512	0.612
SentiStrength	0.728	0.497	0.59	0.666	0.719

The engine outperforms other prevailing engines in correctly identifying negative emotion, as well as two popular engines in identifying positive emotions, and enables better assessment of emotion[4].

3. Data

We used CustSent to explore the presence and dynamics of emotions in three large corpora of data. The data are of customer interactions with service employees in three companies from different industries (telecommunication, retail and air transportation). All interactions were conducted between October and December 2016. The full data set comprises 1.14 million full interactions, and close to 14 million individual text messages (See Table 7 for the amount of messages and interactions by company). These data offer comprehensive documentation of customer emotions from different perspectives, as we detail in Section 5.

Table 7 Full data analyzed in the study by company.

Company	Mean number of customer messages in interaction (SD)	Number of interactions
Transportation	8.12 (7.78)	25,714
Retail	8.32 (7.80)	439,585
Telecommunication	14.8 (13.2)	677,936

The data include the automated CustSent assessments of the presence and intensity of positive and negative emotions in individual customer messages. Messages are identified as *Positive* when CustSent assigned a score higher than 0, as *Negative* when CustSent assigned a score lower than 0; if CustSent assigned a score of 0 the message is identified as *No emotion*[5]. From the message

[4] Note here that all the data we analyzed did not include any identifying information. Customer privacy and anonymity was fully maintained.

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

level scores we can aggregate to the full interaction-level. We define interactions as *Positive* if they include at least one positive message and no negative messages, and define interactions as *Negative* if they include at least one negative message and no positive messages. Interactions are defined as *Multiple emotion*, if they include at least one positive and one negative message, and as *No emotion*, if CustSent assigns a score of zero to all the messages in the interaction.

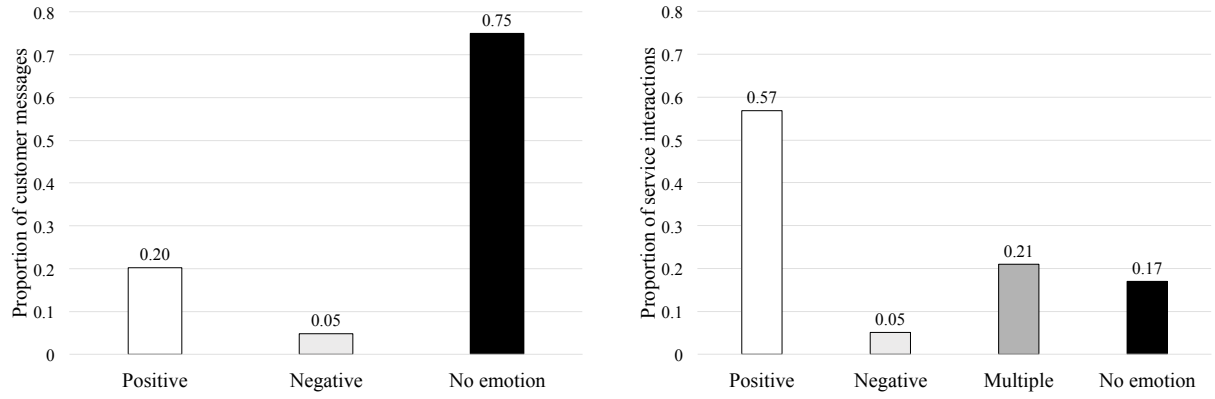
In the remainder of this paper, we use these classifications to report various insights about customer emotions from different perspectives. We report on multiple analyses, some of the individual data sets (companies) separately, and some on all the data combined, depending on the specific research question. Combining all the data would potentially create unintended biases in some of the analyses, due to the different volume of the three companies. For reasons of brevity we do not report all the analyses of the three companies. Instead, we report analyses of one of the companies, randomly selecting the company reported in each analysis. Unless otherwise explicitly stated, all the results we report are tested using the relevant statistics, and found to be statistically significant ($p < 0.05$). For all the figures and analyses in the next section, we denote n as the sample size.

4. Findings

4.1. Type and frequency of sentiment in service interactions

A first broad insight of our analyses is that most customer messages do not contain emotions (Figure 3a). However, when looking at complete service interactions, more than 80% of interactions contain at least one expression of an emotion (Figure 3b). These findings support the assertions that emotions prevail in service interactions (Berry 1999, Rafaeli and Sutton 1987), which clearly implies that management should consider customer emotion in planning and execution of service operations. Interestingly, contrary to common belief the dominant emotion is positive emotion. Pure negative emotions are expressed by only a small proportion of customers ($< 10\%$), and appear in 5% of customer messages, while positive emotion appears four times more than negative emotion.

A more refined analysis, of the intensity of customer emotions, is reported in Figure 4. This analysis concentrates only on interactions in which some customer sentiment was detected. We sum up all (positive and negative) sentiment intensities in all customer messages within each interaction



(a) Emotion in Messages $[n = 3,659,053]$ (b) Emotion in Interactions $[n = 439,585]$

Figure 3 Frequencies of emotion in customer messages and in full service interactions [Retail, 10-12/2016]

in which emotions are present. We report this analysis for transportation interactions ($n = 21,122$), and find that in a quarter of the interactions (27%), positive and negative emotion intensities are equally present (evident in their sum being zero). Figure 4 again shows that positive emotion is more prevalent than negative emotion, even when we look at overall emotion in an interaction.

We also observe that the frequency of high intensity positive emotion is higher than the frequency of high intensity negative emotion. The tail of the summed positive emotions is significantly heavier than the tail of the summed negative emotions. Empirically, for example, 5% of the interactions have a sum of intensity lower than -2, while the comparable index with positive emotion is substantially higher, with 21% of interactions having a sum of intensity higher than 2. Thus overall, customers who express emotion in service interactions are more likely to express positive emotions, and likely to express them more intensely than negative emotion.

The large amount of data we analyzed allows to compare emotions expressed in customer interactions from different companies. We analyzed data from three companies (telecommunication, retail and transportation), which we selected because they represent distinct types of services (Wirtz and Lovelock 2016). As evident in Figure 5, less than 20% of the interactions in all companies contain no emotion. And, in the three companies, 38% or more of the interactions contain positive emotion, while 10% or less contain negative emotion. Hence, the general pattern of emotion distribution is similar in the three companies. Notwithstanding, a test of independence (Chi Square) shows a statistically significant difference between the distributions of emotions between companies.

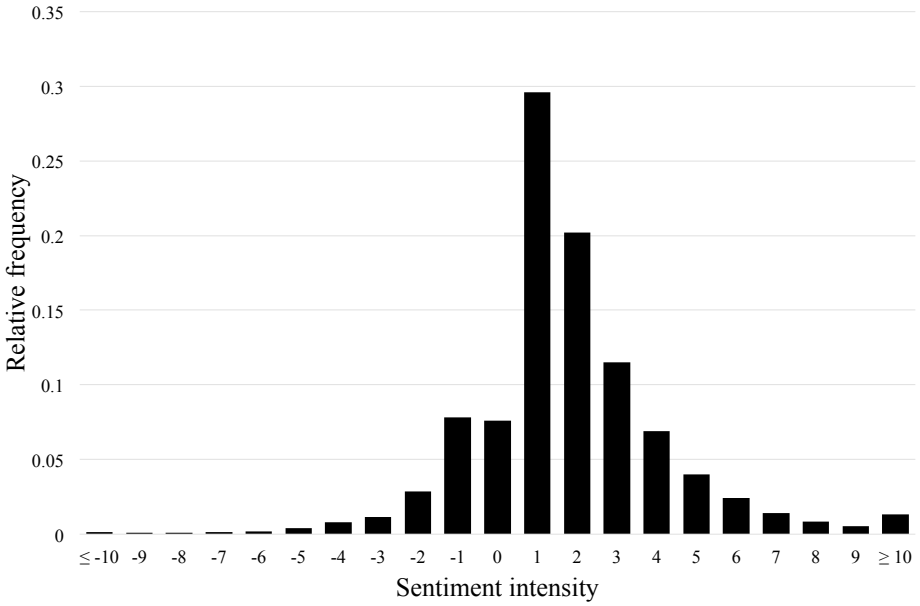


Figure 4 Distribution of summed intensity of customer positive and negative emotions [Transportation, 10-12/2016, $n = 21,122$**]**

A comparison between companies highlights that the telecommunication company has substantially fewer interactions that include only positive emotions than the other two companies (38% as compared to 52% in transportation and 57% in retail), and more negative emotions (10% vs. 7% in transportation and 5% in retail). Moreover, the telecommunication company has the highest proportion of interactions with *Multiple emotion*. In the next section, we will show an additional difference between companies in the dynamics of the interaction.

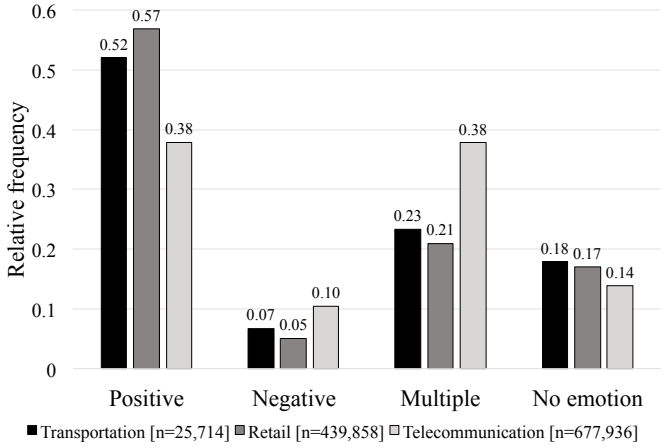


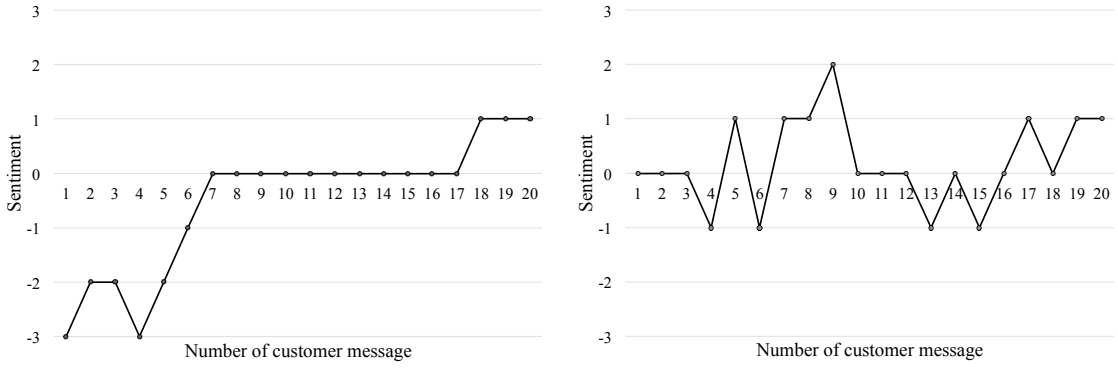
Figure 5 Sentiment frequency in service interactions in three companies [10-12/2016]

4.2. The flow of sentiment within service interactions

The CustSent model further avails a look at the unfolding of customer emotion over the course of service interactions. In other words, the model can depict the flow of the sequence of emotions that a customer expresses throughout a service interaction, or the emotional journey the customer travels. Since different customers bring different needs, problems, and expectations into the service interactions, their emotion flows throughout the interaction are likely to be different. To illustrate, Figure 6 presents two flows of sentiment in the interactions of two specific customers. Both interactions are with employees of the same telecommunication company and both contain exactly 20 customer messages, yet, as evident in the figure, the customers emotional journeys differ.

The visual depiction of the flow of emotion in Figure 6 highlights emotional characteristics of the interactions. The first trimester of the interaction in Figure 6a is negative, indicating that the customer entered the interaction with negative emotions. The second trimester is neutral, so the initial negative emotions with which the customer started the interaction are less central. The interaction ends with the customer expressing positive emotions, suggesting that whatever caused the negative emotions was resolved. In contrast, the customer in Figure 6b seems to be on an emotional roller coaster, with no clear pattern to the emotions; the customer wavers between expressing positive and negative emotions in different parts of the interaction. Figure 6 suggests that service employees must handle very different emotional dynamics when handling multiple customer interactions (Rafaeli and Sutton 1987).

Beyond the anecdotal story presented by Figure 6, aggregation of data that is used to create Figure 6 across all customers can uncover patterns of emotions within service interactions in general. For example, Figure 7 compares the sentiment in all first customer messages in an interaction to the sentiment in all last messages (gray vs. black lines in Figure 7). As evident, emotions are more positive at the end than at the beginning of service interactions. Also evident in Figure 7, is that this effect is not strongly related to time of the day in which the interaction occurs. There is a small effect of the first and last hour of the day in the emotion in the end of the interaction. The



(a) Interaction 1 (b) Interaction2

Figure 6 Flow of sentiment throughout two customer interactions [Telecommunication]

beginning of the day has slightly more positive emotion (comparing 08:00 to 09:00), and the end of the day has slightly lower positive emotion (comparing 21:00 to 22:00). This could be because customers are less patient at the end of the day, or because agents are tired and less effective.

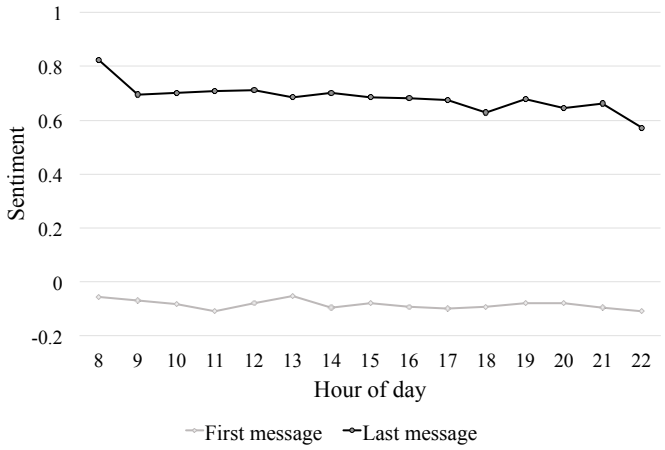


Figure 7 Customer sentiment in the beginning and the end of a service interaction at different hours of the day [Transportation, 10-12/2016, $n = 25,668$]

A deeper understanding of typical emotion dynamics in service interactions is obtained by aggregating the flow of sentiment of single interactions (Figure 6) across all interactions. Since the length of interactions varies, such aggregation and comparison requires standardizing the length of interactions. We standardize the length of interactions by splitting each interaction into 10 roughly equal sections, such that sections in different interactions may comprise a different number of messages,

but all interactions comprise exactly 10 sections (or 10 quartiles). We then average the sentiment of all messages in each section, and thus define each interaction as comprising 10 sections, and 10 sentiment scores. This standardization allows us to aggregate the emotion in similar sections of interactions (e.g., beginning, middle or end of the interaction), and to capture and describe the flow of sentiment over the course of multiple interactions[6].

More precise details of this standardization approach are provided next. For a conversation C with n customers messages and sentiment scores $\{s_i\}_{i=1..n}$, we can split the conversation to a smaller number of sections $m \leq n$ of roughly equal length (between $\lfloor \frac{n}{m} \rfloor$ and $\lceil \frac{n}{m} \rceil$). The intervals that define the different sections are $\{[j_{i-1}+1, j_i]\}_{i=1..m}$ where the boundaries are defined as the closest integers to the $\frac{n}{m}$ split, i.e. $j_i = \lfloor \frac{n}{m} \times i + 0.5 \rfloor$. For each section, we calculate the average of the sentiments, i.e. for section i the average is $s'_i = \frac{1}{j_i - j_{i-1}} \sum_{j=j_{i-1}+1}^{j_i} s_j$. These averages create the *standardized sentiment flow (SSF)* of size m for conversation C:

$$SSF_m(C) = \{s'_i\}_{i=1..m}.$$

In order to aggregate multiple interactions, we choose a fixed size m and create *SSF*s, one for each interaction. This means that for k interactions we get k standardized sentiment flows $\{SSF_m(C_i)\}_{i=1..k}$. We then average each of the m points across these *SSF*s to create an *aggregated sentiment flow* (ASF). For the j-th points in the SSFs of the k interactions denoted by $\{s'_{i,j}\}_{i=1..k}$ the average is $s''_j = \frac{1}{k} \sum_{i=1}^{k} s'_{i,j}$ and thus the aggregated sentiment flow of these k interactions is $ASF_m(\{C_i\}_{i=1..k}) = \{s''_j\}_{j=1..m}$.

Figure 8 presents the aggregated sentiment flow of the population of interactions of the telecommunication company, that have at least 10 customer messages. It suggests three stages within a

[6] We report this only for the subset of interactions that included 10 or more customer messages. We repeated this analysis on all of the data, including shorter interactions, as a robustness check. In the repeated analysis we stretched short interactions, that include 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.

standard interaction; An opening, which includes negative emotions; a middle, with mostly neutral messages; and an ending, with more positive emotions. Negative emotion at the start of interactions is not surprising, since customers typically initiate a service interaction when they have a problem; as the interaction evolves, emotions are less likely because the focus is on resolving the problem at hand; toward the end of the interaction, when the problem is close to being resolved, emotions are likely to be more positive. An interesting direction for future research may be to identify the breaking or tipping point of interactions, from which the sentiment begins to improve. Managerially this is an important goal because it can help improve the duration of positive emotions that customers associate with a service procedure.

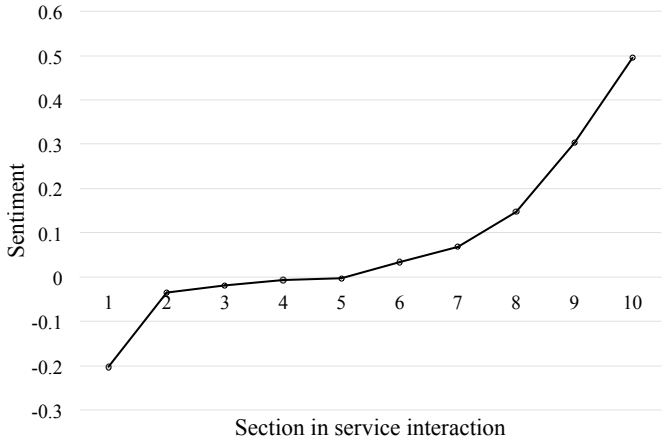


Figure 8 Aggregated sentiment in sections of service interactions [Telecommunication, 10-12/2016, $n = 390,438$]

Extending Figure 8, we compare sentiment flows in the three companies, depicted in Figure 9. This comparison again shows generally less positive emotions in the telecommunication company *throughout* the whole interaction. We cannot explain these findings, though they may be indicative of a different service culture in the telecommunication company, or they may be indicative of differences between industries, wherein telecommunication attracts more negative customer emotions.

Another analysis is a search for types, or prototypes of customer interactions. We used k-means clustering (MacQueen 1967) to identify such prototypes, and Figure 10 reports results of one

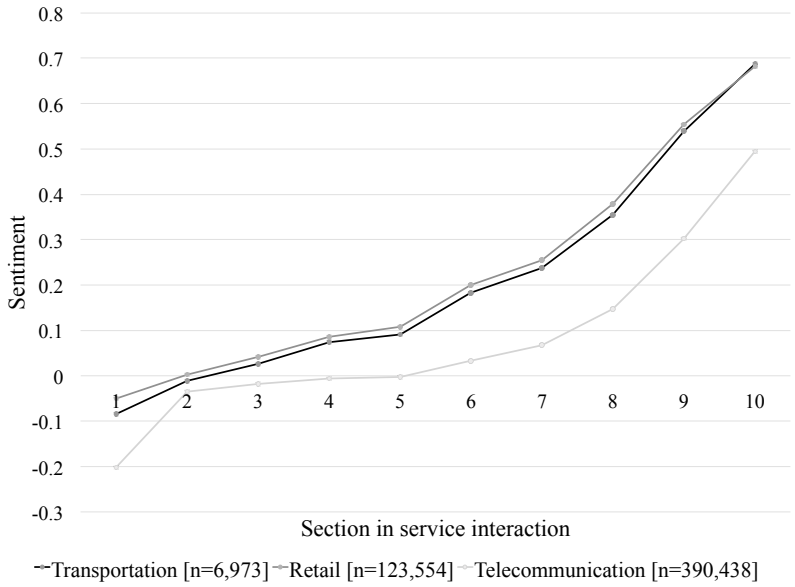


Figure 9 Aggregated sentiment flow in three companies [10-12/2016]

analysis, with four (4) clusters defined[7]. We observe little variation of sentiment in the initial sections, and noticeable variations in the second half of the interaction. The pattern of emotion until the 6th or 7th section is basically flat, with close to zero (0) emotion. The 7th section seems to be a tipping point (cf., Gladwell 2000). This point is the inflection point of the curves, from which the direction of the curve changes, from negative or neutral to positive. These findings remind of a pattern identified by Gersick (1989, 1988) of a "tipping point" in the dynamics of meetings and teams, in which the culmination of the team performance is determined. Service interactions are of dyads, which by definition are "minimal" two-person groups (cf., Forsyth 2009, Brown 1988). And our analyses suggest that these dyads (or mini-groups) also have a tipping point in their dynamics.

The notion of interaction patterns, and of a potential tipping point in service interactions can also be connected to the outcome of service interactions. Different patterns of emotion in a service interaction may relate to service outcomes, as evident in customer satisfaction indices. We explore this in the next section.

[7] We performed this analysis also for 2, 3, and 5 clusters and obtained similar results.

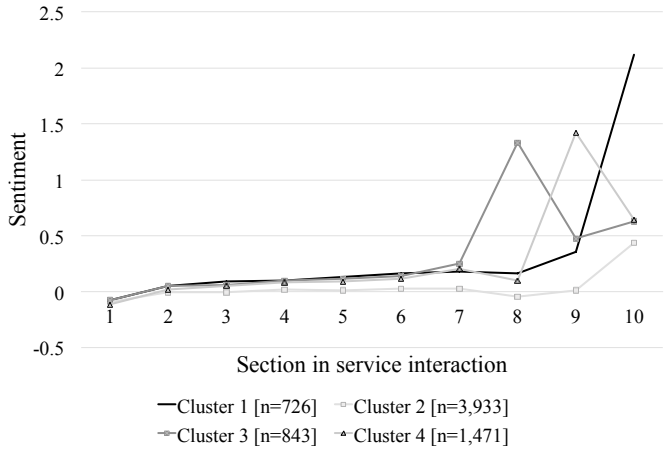


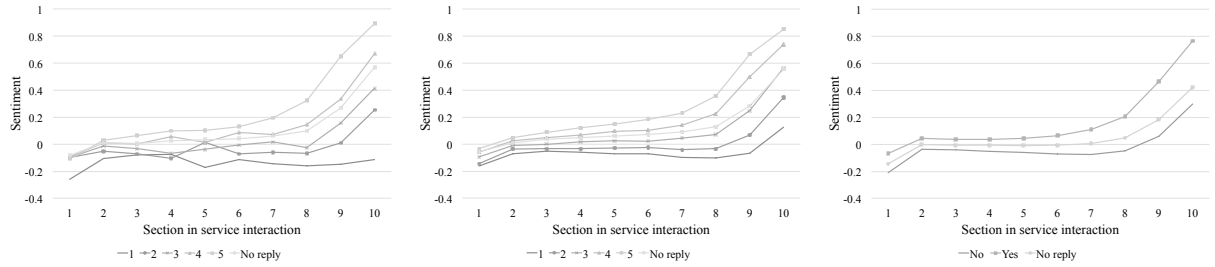
Figure 10 Clusters of emotion patterns in service interactions [Transportation, 10-12/2016, $n = 6,973$]

4.3. Relating customer sentiment to customer satisfaction

Our analyses thus far reported on dynamic assessment of customer emotion as customers express themselves over the course of their service interactions. These assessments reflect customer feelings during the interaction, and a useful analysis is of the relationship between these customer emotions and customer satisfaction after a service interaction has ended. Current assessments of customer satisfaction are available only after a service interaction has ended, in what is known as "post-service surveys". These surveys are known to engender multiple problems; As Kaplan (2016) noted, satisfaction surveys are considered a nuisance, and a turnoff to many customers; responses to surveys often show multiple biases; and people are known to lose patience and often refuse to respond. Our data include satisfaction measures of customers who responded (cf., https://www.liveperson.com/uk/services/technical-support/about-post-chat-survey), and indicate that only approximately 25% of transportation customers, 30% of telecommunication customers, and 47% of retail customers responded. All this suggests that tracking customer feelings during a service interaction may be a more valid indicator of overall customer satisfaction. We report on the relationship of emotions to customer satisfaction below.

To test this idea, we assess the relationship of customer emotion as detected by CustSent to the satisfaction reports in our data. We compare customer emotions to three satisfaction surveys commonly used as service satisfaction measures: (i) Customer evaluation of employee performance,

(ii) "Net Promoter Score" (NPS; Reichheld 2003, and (iii) "First Contact Resolution" (FCR; Hart et al. 2006). We elaborate on these measures below. Figure 11 shows insightful patterns of relationship between emotions and these three satisfaction measures, showing that interactions typically begin with a low negative emotion score, regardless of customers' satisfaction from the service. This is consistent with the idea that customers initiate a service interaction because they have a problem. Figure 11 also shows a pattern of emotion evolution, which unfolds during the interaction, and varies in ways that relate meaningfully to customer satisfaction measures. The emotional patterns stack neatly one above the other, with minimal overlap. A steeper trajectory of improvement of customer emotion from negative to positive emotion indicates a more rapid move from negative to positive emotion, and seems to relate to a higher satisfaction score at the end of the interaction[8]. This occurs in all companies and with the three service indices, as we detail next.



(a) Employee performance [Transportation, $n = 6,973$]

(b) Loyalty (NPS) [Retail, $n = 123,554$]

(c) Customer First Contact Resolution (FCR) [Telecommunication, $n = 390,438$]

Figure 11 Customer sentiment in sections of interaction by customer evaluations

Figure 11a compares customer emotion and ***customer assessment of employee performance***. The question presented to customers may be "Thinking about the employee that you just chatted with, how would you rate them?" and responses presented to the customer are: Excellent, Good, Average, Below average, and Poor. These responses are translated into numbers from 1

[8] For sake of comparison, we also report the pattern of emotions of customer who did not respond to any satisfaction measure. We note that the emotions of non-respondents seem like the average of the customers who did respond.

(poor) to 5 (excellent). Interactions that include a minimal level of negative customer emotion throughout the interaction, tend to end with lower employee assessments of the employee than the overall customer population (11a). In contrast, the more steeply positive the gradient of customer emotions during an interaction, the more positive evaluations of the service employee after the interaction has ended. This suggests that CustSent can also be used for performance evaluation of service employee, and as a mechanism for creating incentive policies. More positive emotion in their latter part of an interaction, is one possible index of better employee performance. Alternatively, one can measure the difference between emotions in initial and final stages of interactions. Such employee evaluation would be automated, objective, unbiased, and complete.

Figure 11b compares customer emotion and a measure of customer long term loyalty, known as the ***"Net Promoter Score"*** (NPS; Reichheld 2003). NPS is assessed through customer responses to one question: "How likely is it that you would recommend our company to a friend or a colleague?" Companies use different response scales, of 0 to 10, where 0 indicates "Not likely to recommend" and 10 indicates "Highly likely to recommend", or a five-point scale, where 1 is "Extremely unlikely" and 5 is "Extremely likely". As evident in Figure 11b, here as well, customers whose overall NPS response indicates low loyalty, maintain generally negative sentiment throughout the interaction. As the NPS evaluation increases, the transition from negative to positive emotion is more rapid, occurring in earlier stages of the interaction, and more steep, leading to a higher level of positive emotion at the end of the interaction. Note that while companies usually summarize the NPS score into three segments and managers often consider only the 9 or 10 scores as positive, our analysis suggests a broader difference in terms of how customers feel during a service interaction.

A final, overall customer assessment called ***"First Contact Resolution"*** (FCR; Hart et al. 2006) is a popular indicator in the service industry, assessed through customer responses to the question "Was your query resolved in this interaction?" (Yes/No). Affirmative responses to the FCR measure occur when the emotion was more positive throughout the interaction, and in particular when the latter part of the interaction included a steep climb in positive emotion (Figure 11c).

More broadly, the emotional patterns in the different graphs demonstrate that there is greater granularity in sentiment customer emotion than any customer survey can depict. Organizations frequently aggregate customer satisfaction reports into bins of high (e.g., responses of 4–5 in a 5 point Likert survey) and low (responses of 1–2). Our sentiment analyses show a more complete picture of the different patterns of evolution of customer emotion that eventually coalesce into distinct satisfaction scores.These graphs again show our observation of a tipping point in service interactions (segment 5 or 6), from which customer emotions stop being negative or neutral, and begin to be more positive. The same point is also where different emotion patterns for different eventual customer evaluations begin to form. The distinction between satisfied and not satisfied customers begins at this point, and is evident in the emotion in the latter part of the interaction.

4.4. An employee perspective: Temporal sentiment dynamics

Customer emotions are experienced not only by the customers, but also by the service employees who interact with the customers, and who can influence their experience. Our data enable the exploration of the employees experience of customer emotions. We begin by exploring the customer emotions experienced by a random customer service employee. Then, we will discuss some alternative forms of aggregating customer emotion across multiple employees. Figure 12 presents the customer emotions experienced by a random employee in the transportation company, during one shift (from 8:00 to 18:00), in a two-minute resolution.

During the course of a work shift employees encounter random "spikes" of customer emotion (Figure 12). The figure shows the total amount of (positive and/or negative) emotion that customers expressed towards the employee as he or she proceeded with the service work. The figure represents a typical employee of a transportation service system. Such employee interacts with anywhere from 22 to 124 (Mean=64; SD=27) customers during a shift. And 99% of interactions vary in duration between less than a minute and 39 minutes (Mean=11.5; SD=9). Each customer interaction has its own unique emotional tone and profile. And agents must move between and constantly adapt to new emotional patterns. This constant move between interactions and recurring adaptation is

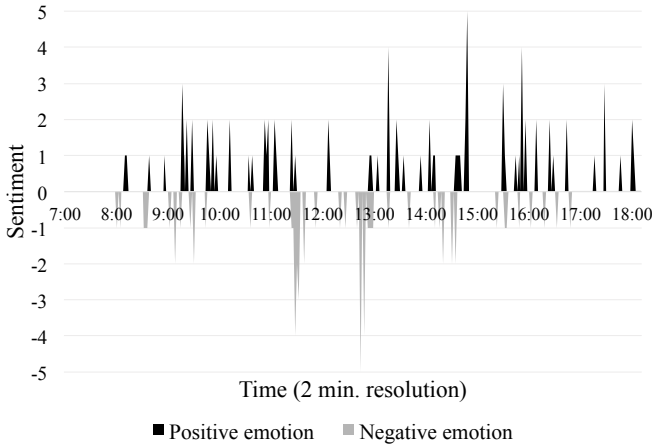


Figure 12 **Customer emotion experienced by one employee during a work day [$n = 124$ customer interactions]**

psychologically taxing (Rafaeli and Sutton 1991), and may explain the high burnout and turnover of service employees (Holman et al. 2007).

Note that there are times where employees do not have to handle a lot of customer emotion (i.e., "emotionally slow"), and other times in which they have to handle extreme customer emotions (e.g., 13:00, 15:00). Processing, handling and reacting to customer emotion requires employee attention, and this distribution supports theoretical assertions of sporadic "emotional tasks" implied by the Affective Events Theory (Weiss and Cropanzano 1996). To our knowledge, this is a first empirical support of the emotional roller-coaster that service employees routinely handle. And the distribution highlights the importance of accounting for customer emotion in employee training, and providing tools and resources to help employees cope with this roller-coaster.

We suggest that routing and staffing strategies should be improved by integrating customer emotion into the considerations. In Yom-Tov et al. (2017b) the authors define emotional load as part of the overall load service employee manage. Related empirical research suggests that such load impacts employee efficiency (Altman 2017) and employee tendency to take unscheduled breaks (Ashtar 2017). Rafaeli et al. (2012) and Miron-Spektor et al. (2011) show that encounters with customers who express negative emotion hamper employee problem solving and cognitive functioning. Therefore, encounters with negative customer emotions require employee recovery, so a short break after a negative customer would provide time for such recovery. Routing policies

can be the tool by which such recovery time can be provided to employees, and accounted for by service management systems. Routing procedures can be developed to take into account the extent of negative customer emotions that an employee recently handled. Algorithms that decide about the allocation of incoming customers can benefit from the inclusion of emotional load analyses in addition to other parameters. Aggregate analyses of the total pattern of negative emotion that an employee is likely to handle over the course of a work shift can further help staffing decisions by accounting for the total amount of recovery time that employees may need. This may impact staffing policies too. Yom-Tov and Rafaeli (2017) explore such policies.

Figure 13 shows that episodes of positive and negative emotion are distributed in a relatively stable pattern over the hours of the day, thus emphasizing that emotional episodes are not prone to temporal effects. The emotional characteristics of service interactions are not different between morning and afternoon hours, for example. This is also evident in Figure 14, in which we use the methods presented earlier for sectioning interactions into a unique structure of 10 sections, to compare the emotional patterns across hours of the day (Figure 14a) and days of week (Figure 14b). The different lines in each figure almost completely overlap. These findings join earlier findings that do not support the presumed phenomena of "Blue Monday" effects (cf., Stone et al. 1985). We clearly see that such notion is not evident in customer emotional expression in service chats.

5. Discussion

Our paper introduces a new approach to studying customer emotions in service interactions. A home grown and validated engine for automated assessment of customer emotions allows us to explore emotions in large samples of genuine customer interactions. The benefits of the new methodology that we present are substantial; it offers a view of customer emotions that is objective, unobtrusive, and fully and directly connected to customers actual expressions and behaviors (Webb et al. 1966). It also identifies emotion dynamics that could not be shown in previous research by overcoming the limitations of previous methods for assessing customer emotions.

Our contribution is therefore threefold: (a) methodological, proposing automated sentiment analysis as a central tool for research and managerial goals; (b) breadth, documenting insights about

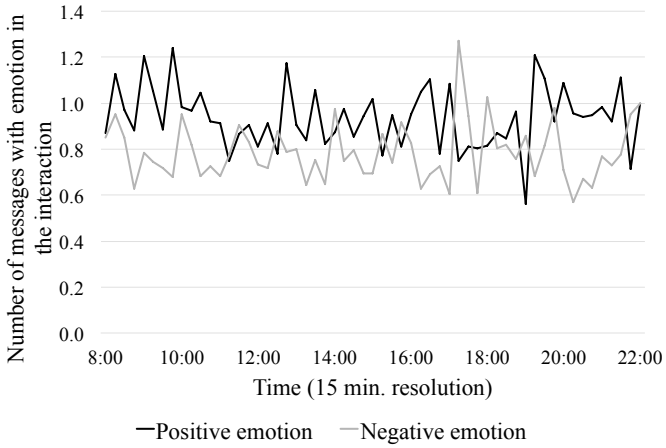
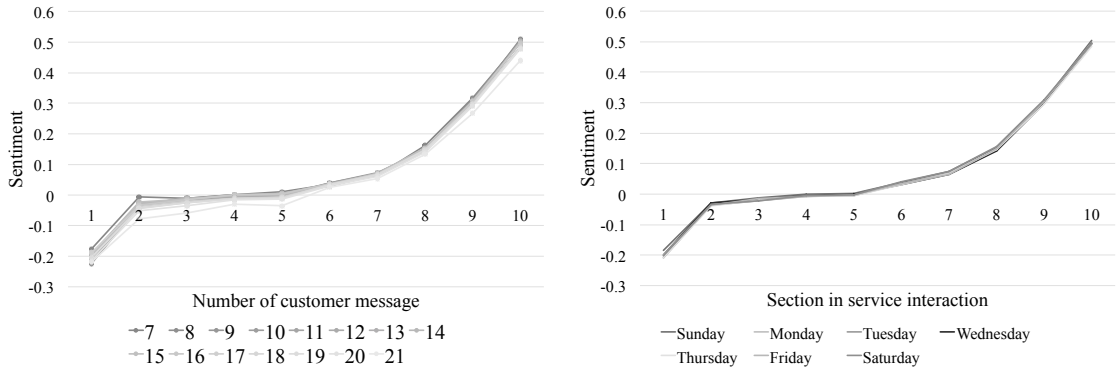


Figure 13 **Frequency of cumulative positive and negative emotion (of all customers) during a work day [Transportation, 10/2016, $n = 11,591$]**



(a) By hour of day

(b) By hour of week

Figure 14 **Aggregated sentiment flow by hour and day [Telecommunication, 10-12/2016, $n = 390,438$]**

dynamics of customer emotions in large scale samples of real customer interactions; and (c) documenting emotion dynamics within customer interactions. The paper also makes an important practical contribution, suggesting new ways for organizations to identify emotions and leverage automated sentiment analysis in service operations. The automated analysis of emotion CustSent offers paves the way for a wide range of analyses of all sorts and forms of big-data, which can promote research and management of operations (George et al. 2016), human resource management (McAbee et al. 2016), and service delivery (Rafaeli et al. 2016, Grewal et al. 2017). Our analyses begin to insinuate the types of issues and questions that future research can address following the procedures that we propose. The amount of relevant data, its velocity (the pace with which it

is accruing and increasing), and its variety (relevant data comes in many forms, and from many sources both within and outside of the organization) coalesce to a rich and exciting research agenda.

The connections that our results described (§4), are enlightening. They propose an emotional pattern to customer service interactions, with a somewhat negative start, that segways into a task-focused and non-emotional plateau. When the service interaction goes well, customer emotions become positive toward the its interaction. This suggests that automated emotion assessment conducted in real-time *during* a service interaction can be exploited to predict customer satisfaction after the interaction. Conversely, such analyses can be used to identify instances where service operations have failed, by tracing customers whose emotions do not traverse to the desirable positive end. Identifying service failures is critical (Tax and Brown 1998). Yet, available research on service failures relies on customer reports and responses to surveys, which come at a delay, and with limited response rates (cf., Casidy and Shin 2015, Joireman et al. 2013, Smith and Bolton 1998). We envision a time in which post-hoc customer satisfaction scores will be replaced by objective measures of customer emotion during the interaction. Such real-time monitoring can impact managerial decisions or policies for supervisor interventions in an unsuccessful interaction, as well as employee assessment and system level control of operational decisions within a service center.

Tools such as CustSent can be used also to investigate effects of customer emotion on employee performance. In other research projects connections of customer emotions and employee work performance are explored; preliminary results indicate that customer sentiment influences employee response time and employee tendency to take unscheduled breaks (Altman 2017, Ashtar 2017). Future research should also consider analyses of employee expressions of emotion. Our initial observations of employee messages show lower variance in their emotions, suggesting that employees regulate their expressions of emotions towards customers. This was referred to as “ “emotional labor” by Rafaeli and Sutton (1987). Therefore, CustSent requires adjustments to correctly support employee emotion detection. Using both CustSent and a designated tool for employee emotion detection will allow a more comprehensive modeling of emotion dynamics in customer service interactions.

Finally, our analysis examines customer emotions in a context that maintains the historical structure, wherein service interactions occur in a customer-employee dyad. The development of the Internet and social media provide access of customer expressions (e.g. anger, satisfaction) to other customers as well. Customers can observe emotions expressed by other customers, in Twitter and Facebook service interactions (see for example Spivak (2017)). Social media service means that customers can incorporate the experience of other customers when they choose and form opinions about a company or a service. Detection of customer emotions in customer reviews has been a central means for automated retrieval and analysis of emotional information. We aim to make the same contribution to the assessment and understanding of customer emotions within service centers. Our CustSent engine, and the view of customer emotions as a dynamic entity integral to service interaction, opens up numerous opportunities to further understand connections between operational decisions, customer satisfaction and employee behavior.

Acknowledgments

We wish to thank Naama Tepper and Shlomo Lahav for initiating the collaboration between the Technion research team and LivePerson. We also thank the following students for helping us in coding the data for testing and validating the sentiment analysis tool: Galia Bar, David Spivak, Gabby Mayer, Cassidy Laidlaw, Laura Blumenfeld, Beaux Ballard.

References

Altman D (2017) *Modeling employee behavioral reactions to emotions expressed by customers: A non-obtrusive examination of customer service employee behavior*. Master's thesis, Technion—Israel Institute of Technology, URL http://www.graduate.technion.ac.il/Theses/Abstracts.asp?Id=30517.

Amabile TM, Schatzel EA, Moneta GB, Kramer SJ (2004) Leader behaviors and the work environment for creativity: Perceived leader support. *Leadership Quarterly* 15(1):5–32.

Ashtar S (2017) *The effect of customer emotion and work demands on employee unscheduled breaks: An investigation of chat-based customer service*. Master's thesis, Technion—Israel Institute of Technology, URL http://www.graduate.technion.ac.il/Theses/Abstracts.asp?Id=30556.

Berry LL (1999) *Discovering the soul of service: The nine drivers of sustainable business success* (New York: Simon and Schuster).

Boiy E, Hens P, Deschacht K, Moens MF (2007) Automatic sentiment analysis in on-line text. *ELPUB*, 349.

Bommannavar P, Kolcz A, Rajaraman A (2014) Recall estimation for rare topic retrieval from large corpuses. *Big Data (Big Data), 2014 IEEE International Conference on*, 825–834 (IEEE).

Brown R (1988) *Group processes: Dynamics within and between groups* (Basil Blackwell).

Buechel S, Hahn U (2017) EMOBANK : Studying the impact of annotation perspective and representation format on dimensional emotion analysis. *EACL 2017*, volume 2, 578–585.

Casidy R, Shin H (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* 27:103–112.

Dasborough MT (2006) Cognitive asymmetry in employee emotional reactions to leadership behaviors. *Leadership Quarterly* 17(2):163–178.

Donaldson SI, Grant-Vallone EJ (2002) Understanding self-report bias in organizational behavior research. *Journal of Business and Psychology* 17(2):245–260.

Elfenbein HA (2007) Emotion in organizations: A review and theoretical integration. *Academy of Management Annals* 1(1):315–386, ISSN 1941-6520.

Forsyth DR (2009) *Group dynamics* (Wadsworth Cengage Learning).

Froehle CM, Roth AV (2004) New measurement scales for evaluating perceptions of the technology-mediated customer service experience. *Journal of Operations Management* 22(1):1–21.

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: tutorial, review, and research prospects. *Manufacturing and Service Operations Management* 5(2):79–141.

George G, Osinga EC, Lavie D, Scott BA (2016) Big data and data science methods for management research. *Academy of Management Journal* 59(5):1493–1507.

Gersick C (1988) Time and transition in work teams : Toward a new model. *Academy of Management journal* 31(1):9–41.

Gersick C (1989) Marking time: Predictable transitions in task groups. *Academy of Management Journal* 32(2):274–309.

Gladwell M (2000) *The tipping point: How little things can make a big difference* (Little, Brown and Company).

Grewal D, Roggeveen AL, Nordfält J (2017) The future of retailing. *Journal of Retailing* 93(1):1–6.

Groth M, Grandey A (2012) From bad to worse: Negative exchange spirals in employeecustomer service interactions. *Organizational Psychology Review* 2(3):208–233.

Hart M, Fichtner B, Fjalestad E, Langley S (2006) Contact centre performance: In pursuit of first call resolution. *Management Dynamics* 15(4):17–28.

Holman D, Batt R, Holtgrewe U (2007) The global call center report: International perspectives on management and employment. Technical report, Cornell University ILR School.

Howard GS, Dailey PR (1979) Response-shift bias: A source of contamination of self-report measures. *Journal of Applied Psychology* 64(2):144–150.

Joireman J, Grégoire Y, Devezer B, Tripp TM (2013) When do customers offer firms a "second chance" following a double deviation? The impact of inferred firm motives on customer revenge and reconciliation. *Journal of Retailing* 89(3):315–337.

Kaplan J (2016) The inventor of customer satisfaction surveys is sick of them, too.

Lakoff G (1984) Performative subordinate clauses. *Meeting of the Berkeley Ling' Society*, 472–480.

Lakoff R (1972) Language in context. *Language* 48(4):907–927.

Larsson R (1993) Case survey methodology: Quantitative analysis of patterns across case studies. *Academy of Management Journal* 36(6):1515–1546.

MacQueen J (1967) Some methods for classification and analysis of multivariate observations. *The fifth Berkeley symposium on mathematical statistics and probability*, volume 1, 281–297 (Oakland, CA, USA.).

Manning CD, Raghavan P, Schütze H (2008) *Introduction to information retrieval*.

Massad N, Heckman R, Crowston K (2006) Customer satisfaction with electronic service encounters. *International Journal of Electronic Commerce* 10(4):73–104, ISSN -.

McAbee ST, Landis RS, Burke MI (2016) Inductive reasoning: The promise of big data. *Human Resource Management Review* 27(2):277–290.

McColl-Kennedy J, Smith A (2006) Customer emotions in service failure and recovery encounters. Zerbe W, Ashkanasy N, Haertel E, eds., *Individual and organizational perspectives on emotion management and display*, chapter 10, 237–268 (Bingley, UK: Emerald Group Publishing Ltd).

McColl-Kennedy JR, Patterson PG, Smith AK, Brady MK (2009) Customer rage episodes: Emotions, expressions and behaviors. *Journal of Retailing* 85(2):222–237.

Miron-Spektor E, Efrat-Treister D, Rafaeli A, Schwarz-Cohen O (2011) Others' anger makes people work harder not smarter: The effect of observing anger and sarcasm on creative and analytic thinking. *Journal of Applied Psychology* 96(5):1065–1075.

Nakov P, Ritter A, Rosenthal S, Sebastiani F, Stoyanov V (2016) SemEval-2016 task 4: Sentiment analysis in twitter. *Proceedings of SemEval*, 1–18.

Narayanan L, Menon S, Spector PE, Journal S, Jan N (1999) Stress in the workplace : A comparison of Gender and Occupations. *Journal of Organizational Behavior* 20(1):63–73.

Oliver RL, Rust RT, Varki S (1997) Customer delight: Foundations, findings, and managerial insight. *Journal of retailing* 73(3):311–336.

Pang B, Lee L (2004) A sentimental education: Sentiment analysis using subjectivity summation based on minimum cuts. *The 42nd Annual Meeting on Association for Computational Linguistics*, 271.

Pang B, Lee L (2008) Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2(1-2):1–135.

Paulhus D, Vazire S (2007) The self-report method. Robins RW, Fraley RC, Krueger RF, eds., *Handbook of research methods in personality psychology*, volume 1, chapter 13, 224–239 (New York: Guilford).

Powers DM (2011) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies* 2(1):37–63.

Pugh SD (2001) Service with a smile: Emotional contagion in the service encounter. *Academy of Management Journal* 44(5):1018–1027.

Rafaeli A, Altman D, Gremler DD, Huang M, Grewal D, Iyer B, Parasuraman A, Ruyter K (2016) The future of frontline research: Invited commentaries. *Journal of Service Research* URL http://jsr.sagepub.com/cgi/content/short/1094670516679275v1.

Rafaeli A, Erez A, Ravid S, Derfler-Rozin R, Efrat-Treister D, Scheyer R (2012) When customers exhibit verbal aggression, employees pay cognitive costs. *Journal of Applied Psychology* 97(5):931–950.

Rafaeli A, Sutton RI (1987) Expression of emotion as part of the work role. *Academy of Management review* 12(1):23–37.

Rafaeli A, Sutton RI (1990) Busy stores and demanding customers: How do they affect the display of positive emotion? *Academy of Management Journal* 33(3):623–637.

Rafaeli A, Sutton RI (1991) Emotional contrast strategies as means of social influence : Lessons from criminal interrogators and bill collectors. *Academy of Management Journal* 34(4):749–775.

Reichheld FF (2003) The one number tou need to grow. *Harvard business review* 81(12):46–55.

Schneider B (1990) *Organizational climate and culture* (Pfeiffer).

Schneider B, Bowen DE (1999) Understanding customer delight and outrage. *MIT Sloan Management Review* 41(1):35–44.

Schumann JH, Wünderlich NV, Wangenheim F (2012) Technology mediation in service delivery: A new typology and an agenda for managers and academics. *Technovation* 32(2):133–143.

Smith AK, Bolton RN (1998) An experimental investigation of customer reactions to service failure and recovery encounters paradox or peril? *Journal of service research* 1(1):65–81.

Socher R, Perelygin A, Wu JY, Chuang J, Manning CD, Ng AY, Potts C (2013) Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the conference on empirical methods in natural language processing (EMNLP)* (Citeseer).

Spivak D (2017) *The effect of emotion exchanges on customer satisfaction in online text-based customer service*. Master's thesis, Technion—Israel Institute of Technology, URL http://www.graduate.technion.ac.il/Theses/Abstracts.asp?Id=30551.

Stone AA, Hedges SM, Neale JM, Satin MS (1985) Prospective and cross-sectional mood reports offer no evidence of a "Blue Monday" phenomenon. *Journal of Personality and Social Psychology* 49(1):129–134.

Strapparava C, Mihalcea R (2007) Semeval-2007 task 14: Affective text. *Proceedings of the 4th International Workshop on Semantic Evaluations*, 70–74 (Association for Computational Linguistics).

Sutton RI, Rafaeli A (1988) Untangling the relationships between displayed emotions and organizational sales: The case of convenience stores. *Academy of Management Journal* 31(3):461–487.

Taboada M, Brooke J, Tofiloski M, Voll K, Stede M (2011) Lexicon-based methods for sentiment analysis. *Computational Linguistics* 37(2):267–307.

Tausczik YR, Pennebaker JW (2010) The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29(1):24–54.

Tax SS, Brown SW (1998) Recovering and learning from service failure. *Management Review* 40(1):75–88.

Thelwall M (2013) Heart and soul: Sentiment strength detection in the social web with sentistrength. *Proceedings of the CyberEmotions* 5:1–14.

Thelwall M, Buckley K, Paltoglou G, Cai D, Kappas A (2010) Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology* 61(12):2544–2558.

van Dolen WM, de Ruyter K (2002) Moderated group chat: An empirical assessment of a new e-service encounter. *International Journal of Service Industry Management* 13(5):496–511.

Webb E, Campbell DT, Schwartz RD (1966) *Unobtrusive measures* (Chicago: Rand McNally).

Wegge J, Van Dick R, Fisher GK, West MA, Dawson JF (2006) A test of basic assumptions of Affective Events Theory (AET) in call centre Work. *British Journal of Management* 17(3):237–254.

Weiss H, Beal D (2005) *Reflections on Affective Events Theory*.

Weiss HM, Cropanzano R (1996) Affective Events Theory: A theoretical discussion of the structure, causes and consequences of affective experiences at work. *Research in Organizational Behavior* 18(1):1–74.

Wirtz J, Lovelock C (2016) *Services Marketing: People, Technology, Strategy* (World Scientific Publishing Co Inc), eighth edi edition.

Wirtz J, Mattila AS (2003) The effects of consumer expertise on evoked set size and service loyalty. *Journal of Services Marketing* 17(7):649–665.

Yom-Tov GB, Rafaeli A (2017) Sentiment-based online routing engine for managing chat-based contact centers. *Manuscript in preparation* .

Yom-Tov GB, Rafaeli A, Altman D, Ashtar S (2017a) Text-based customer service: Using big-data to connect customer emotion to service operations. *POMS 2017*.

Yom-Tov GB, Rafaeli A, Ashtar S, Altman D (2017b) The anatomy of load. *Manuscript in preparation* .

Zhang L, Ghosh R, Dekhil M, Hsu M, Liu B (2011) Combining lexicon-based and learning-based methods for twitter sentiment analysis. Technical report, HPL-2011-89, HP.

Protocol main guidelines

The protocol main guidelines included:

1. Coding of each message reflects chat history up to present message;

2. Coding should take into account explicit emotion (written emotional term) and implicit emotion (subtext);

3. Non-verbal cues (i.e., CAPS LOCK, punctuation marks, emojis, messages length, etc.) amplify emotion intensity;

4. Allow emotion carry-over (i.e., if the previous message was negative and negative emotion continues in the next message within the same interaction, negativity of current message is enhanced).

Table EC.1 Negative emotion rules.

Negative Emotion Feature	Meaning/Examples
Negative expressions	"called X times","ASAP", "your fault"
No pleasantness	Lacking expected politeness/friendliness/agreeableness
Customer complaint/problem	If major/repeated, negative emotion is higher
Aggression	Rudeness, cynicism, sarcasm, refusal
Demanding action	"let me speak to your manager/boss/supervisor"
Mention waiting time	"I have been waiting for/I am still waiting "
Customer threat	Mentions cancellation, court, media
Mentioning the problem again	"The phone has stopped working"

Table EC.2 Positive emotion rules.

Positive Emotion Feature	Meaning/Examples
Positive emotion terms	"great", "awesome", "Im so glad..", "happy", "okay thank you"
Small-talk	off-topic conversation
Inclusion	"we", "lets.." - use of plural
Humor	Makes jokes, use of happy emoticons, responds to jokes
Positivity towards interaction	"It has been my pleasure", "Ive enjoyed our talk" "Thank you for all the information and now I can breathe easy. You have been a wonderful help"
Empathy	Understanding the other persons feelings: "I am sorry to hear that"
Intimacy	Sharing personal information
Politeness	"Thank you!" "Thanks", "May I", "Please", "Sorry, I misspelled", "Can you?" "Maybe I can speak to the manager?", "I wonder if"
Wordiness	"Hi there"