CustSent: A tool for automatic identification of customer emotion in chat services
A technical report

Michael Natapov
michaelna@liveperson.com
LivePerson Inc.

Anat Rafaeli, Galit B. Yom-Tov, Shelly Ashtar
anatr@ie.technion.ac.il, gality@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

---

1. Introduction

We begin by describing a new tool that we developed for automatic detection of emotions in customer messages in chat service interactions. 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. In short, the paper provides first a tool for monitoring emotions within text-based service interactions.

2. 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 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.

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 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. 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.

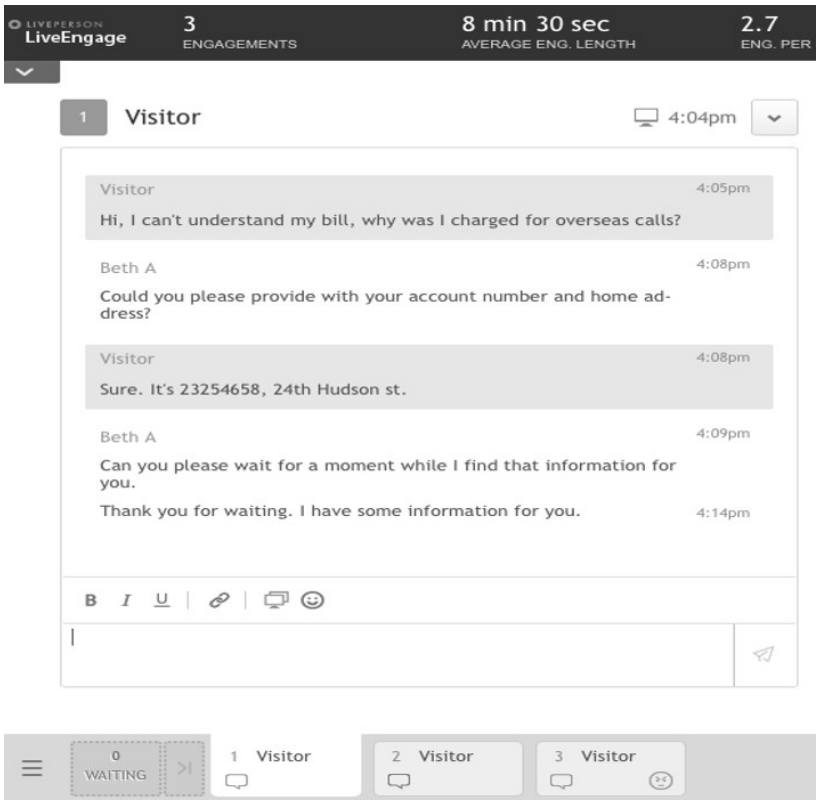


Figure 1 **An example of service interaction between employee and customer**

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.

2.2. 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 §**??**). 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.

2.3. 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)

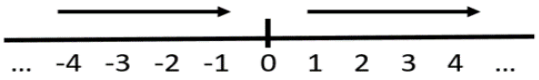
like (0) → *don't like* (-1) → *really don't like* (-2)

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.

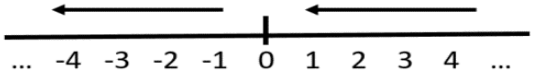
2.3.1. Sentence level rules: Updating emotion scores from anchors to scores for full messages The context-defined scores of anchors in a message are added up to create the preliminary score of a message. For accurate indication of overall emotion in a message, additional analyses of complete sentences 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:

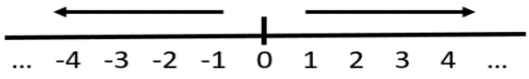
Adding: shifting a score in a positive direction



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

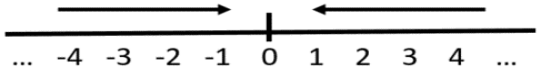


Table 1 **Sentence level rules for modifying the emotion score created by lexical words.**

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)

3. 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.

3.1. 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%.

3.2. 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 ($\sim$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

3.3. 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) 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.

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).

[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.

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

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.

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

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. 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].

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.

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

References

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.

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).

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.

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

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

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.

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.

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

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.

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.

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.

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

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).

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).

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.

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.

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.

Appendix A: 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 7 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 8 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”