RC24754 (W0902-116) February 27, 2009
Computer Science

IBM Research Report

Towards Real-Time Measurement of Customer Satisfaction Using Automatically Generated Call Transcripts

Youngja Park, Stephen C. Gates
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Towards Real-Time Measurement of Customer Satisfaction Using Automatically Generated Call Transcripts

Youngja Park
IBM T. J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598, USA
young_park@us.ibm.com

Stephen C. Gates
IBM T. J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598, USA
scgates@us.ibm.com

ABSTRACT

Customer satisfaction is a very important indicator of how successful a contact center is at providing services to the customers. Contact centers typically conduct a manual survey with a randomly selected group of customers to measure customer satisfaction. Manual customer satisfaction surveys, however, provide limited values due to high cost and the time lapse between the service and the survey.

In this paper, we demonstrate that it is possible to automatically measure customer satisfaction by analyzing call transcripts enabling companies to measure customer satisfaction for every call in near real-time. We have identified various features from multiple knowledge sources indicating prosodic, linguistic and behavioral aspects of the speakers, and built machine learning models that predict the degree of customer satisfaction with high accuracy. The machine learning algorithms used in this work include Decision Tree, Naive Bayes, Logistic Regression and Support Vector Machines (SVMs).

Experiments were conducted for a 5-point satisfaction measurement and a 2-point satisfaction measurement using customer calls to an automotive company. The experimental results show that customer satisfaction can be measured quite accurately both at the end of calls and in the middle of calls. The best performing 5-point satisfaction classification yields an accuracy of 66.09% outperforming the *DominantClass* baseline by 15.16%. The best performing 2-point classification shows an accuracy of 89.42% and outperforms both the *DominantClass* baseline and the *CSRJudgment* baseline by 17.7% and 3.3% respectively. Furthermore, Decision Tree and SVMs achieve higher F-measure than the *CSRJudgment* baseline in identifying both satisfied customers and dissatisfied customers.

Categories and Subject Descriptors

H.4 [**Knowledge Management**]: Mining and representing text, Classification

General Terms

Algorithms, Design, Experimentation, Measurement

Keywords

Customer Satisfaction, Contact Center Calls, Speech Analytics, Natural Language Processing, Text Mining, Classification, Machine Learning

1. INTRODUCTION

Contact centers are critical interfaces between companies and their customers. The top two goals of contact centers are reducing operational costs and improving customer satisfaction, i.e., providing the best quality services at the lowest possible cost. The two goals have been perceived not compatible and having tradeoffs [1]. Companies have mostly focused on achieving the first goal by automating critical processes or outsourcing customer service to other countries with lower labor cost. Most research for contact centers have also been drawn to developing tools for improving agent productivity and saving the costs. Those tools range from real-time agent assistance [22] to automatic call monitoring [38] and semi-automated call logging [4].

Customer satisfaction (C-SAT) is a very important indicator of how successful a contact center is at providing services to the customers, and has been widely used in evaluating the performance of a contact center. Research has shown that customer satisfaction has a strong correlation with profitability [11] and also has strong positive effects on customer retention [27]. A study by *Bain & Company* found that, for many companies, an increase of 5% in customer retention can increase profits by 25% to 95% [28]. However, unlike productivity enhancement and cost saving, it is very hard to objectively measure customer satisfaction.

Most contact centers conduct a manual survey with a small group of customers to measure customer satisfaction. A manual customer satisfaction survey is typically conducted via a telephone interview or a mail-in form, in which customers are asked to evaluate each statement in the questionnaire using a 5-point Likert scale [17]. A typical 5-point question on customer satisfaction is answered as "Completely Dissatisfied", "Somewhat Dissatisfied", "Neutral", "Somewhat Satisfied", or "Completely Satisfied".

Manual customer satisfaction surveys pose three major limitations. First, they are very expensive since most companies hire an external market research firm to conduct a survey. Second, because of the cost, the survey size is typically very small, and, thus, the conclusions drawn from

the survey are not very reliable. Typically, only 1–5% of callers are surveyed, and of these, only a small fraction responds to the survey. A recent study finds that response rates have been falling across all forms of survey research for decades [2]. Third, a manual survey is typically conducted a couple of weeks after a case is finally closed, and, therefore, it is often too late to take an action to prevent customer defection.

Therefore, a tool that can automatically measure customer satisfaction for every call would be highly valuable. Such a tool enables companies to measure customer satisfaction for each and every call. Furthermore, with a real-time speech transcription system, customer satisfaction can be measured in real-time allowing supervisors to take over a call when a customer becomes unhappy and to resolve the customer's issue.

In this work, we present a fully automated method for measuring customer satisfaction by analyzing automatically transcribed calls. The main technical contributions of the work are two folds. First, we identified various features which are highly correlated with C-SAT scores. The features indicate prosodic, linguistic and behavioral aspects of the speakers, and are automatically extracted from call transcripts and information stored in contact centers' database. Second, we developed machine learning models that predict, with high accuracy, customer satisfaction based on the automatically extracted feature set.

Experiments are carried out with 115 customer calls to an automotive company for a 5-point satisfaction measurement (i.e., from "1" to "5") and a 2-point satisfaction measurement (i.e., "satisfied" vs. "dissatisfied") using four widely used machine learning algorithms: Decision Tree, Naive Bayes, Logistic Regression and Support Vector Machines (SVM). Two sets of customer calls are used in the experiments; one comprising the entire conversations, and the other comprising only the first half of conversations.

The performance of automated systems are measured via 10-fold cross validation and are compared with two baseline methods. The first baseline method is an artificial classifier which assigns the dominant class to all calls (a.k.a, *DominantClass*). The second baseline is the customer service representative (CSR)s' judgment on customer satisfaction (a.k.a., *CSRJudgment*).

The experimental results show that customer satisfaction can be measured quite accurately both at the end of calls and in the middle of calls. The best performing 5-point satisfaction classification yields an accuracy of 66.09% outperforming the *DominantClass* baseline by 15.16%. The best performing 2-point classification shows an accuracy of 89.42% and outperforms both the *DominantClass* baseline and the *CSRJudgment* baseline by 17.7% and 3.3% respectively. Furthermore, Decision Tree and SVMs perform better than the *CSRJudgment* baseline in identifying dissatisfied customers achieving 11.5% and 3.2% higher F-measure respectively.

2. COMPARISON WITH RELATED WORK

Customer satisfaction has been said to be one of the most widely studied areas in marketing [3], but there has been little attempt to automatic customer satisfaction measurement. Recently, Godbole and Roy proposed a tool that help contact center Quality Analysts analyze customer feedback text by providing text classification and interactive document labeling [10]. To the best of our knowledge, however, there has been no previous research on customer satisfaction measurement by analyzing automatically generated call transcripts.

Some related bodies of work has been done in the text mining and natural language understanding areas. They include emotion detection in spoken dialogue [18, 8, 32], sentiment analysis and classification [24, 34, 37, 12, 9, 23, 36, 13] and opinion mining [14, 15, 16] for customer review or feedback documents. However, emotion or sentiment detection alone is not sufficient for measuring customer satisfaction. We analyzed contact center calls to study the relationship between customer satisfaction and the use of sentiment words by the customers. Figure 1 depicts the composition of "satisfied" calls and "dissatisfied" calls in terms of the differences in the number of positive sentiment words and the number of negative sentiment words spoken by the customers.

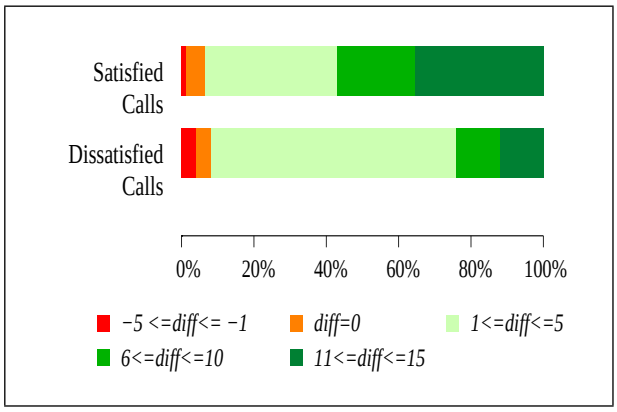

Figure 1: Relationship between customer satisfaction and sentiment words. The chart shows the comparison of "dissatisfied" calls and "satisfied" calls with respect to the relative use of positive sentiment words and negative sentiment words. *diff* is computed by subtracting the number of negative sentiment words from the number of positive sentiment words (i.e., "positive" - "negative") spoken by the customers.

As we can see from the figure, both satisfied customers and dissatisfied customers use more positive sentiment words regardless of their satisfaction level. Only 8% of dissatisfied customers use same number or more negative words than positive words, while 6.4% of satisfied customers also used same or more negative words than positive words. The analysis result indicates that the difference between the positive sentiment words and the negative sentiment words spoken by customers in "satisfied" calls and "dissatisfied" calls is negligible. This analysis results motivate us to look beyond customers' sentiment for measuring customer satisfaction.

The main differences of our work from the related work are the following. Firstly, customer satisfaction is an overall judgment based on cumulative experience with the service and is influenced by multiple factors, including the customer service quality, the time duration spent to have the issue resolved, whether a compensation (or other goodwill token,

e.g., discount or reimbursement) was offered, to name a few. Therefore, to capture the influence of these multiple factors, various knowledge sources need to be exploited to estimate the level of customer satisfaction. In this work, we identified both structured and unstructured features which are highly correlated with C-SAT scores.

Secondly, unlike review or feedback text which are intended to express the authors' opinions, customer calls often contain no explicit emotional expressions or multiple emotional states. Some customers do not express their sentiment or satisfaction level explicitly during a call. Some customers change their sentiment as the call progresses and the issue gets resolved. Some customers expresses different sentiments toward different objects in a call. In the automotive company's case, many customers express their dissatisfaction with the dealership, but they are generally satisfied with the contact center service.

Thirdly, automatic call transcripts are highly noisy and fragmentary due to word recognition errors of the automatic speech recognition (ASR) system and high rate of interruptions and repeats during conversations. Therefore, applying text mining on automatic call transcripts is much more challenging than on review-type text.

Lastly, most of the related work focused on the binary distinction of positive vs. negative for an opinionated text. Pang *et al.* attempted to generalize the problem of categorizing opinionated text into a finer-grained classification task (three or four classes) [23]. In this work, we conduct experiments for both a binary and a 5-ary distinction of customer satisfaction.

3. PROBLEM DESCRIPTION

Customer satisfaction has traditionally been measured by interviewing a small set of selected customers. C-SAT surveys often measure customer satisfaction level from "1" to "5" using a 5-point Likert scale. However, the differences among the scores are very hard to distinguish even for humans. Especially, the distinctions between "1" ("completely dissatisfied") and "2" ("somewhat dissatisfied"), and between "4" ("somewhat satisfied") and "5" ("completely satisfied") are very vague.

The main goal of conducting customer satisfaction survey is in identifying satisfied customers and dissatisfied customers to evaluate the performance of their contact center and to identify areas for service quality enhancement. Therefore, in most cases, a binary classification of customers into satisfied customers and dissatisfied customers might be sufficient.

In this work, we investigate the feasibilities of real-time measurement of customer satisfaction for both classification scenarios.

1. 5-point satisfaction classification assigning contact center calls into five C-SAT score groups
2. 2-point satisfaction classification assigning contact center calls into "satisfied" and "dissatisfied" categories

The main goals for this study are two-fold. First, we aim to identify feature combinations that are highly correlated with customer satisfaction scores and can be automatically extracted from data sources available in most contact centers. Second, we aim to identify machine learning approaches which can measure the degree of customer satisfaction with reasonably high accuracy.

4. THE APPROACH

In this section, we describe the four machine learning algorithms used in this work, and explain the features in great detail.

4.1 Learning Methods

To our knowledge, this is the first attempt for applying natural language processing (NLP) and machine learning technologies to automatically measure customer satisfaction by analyzing call transcripts. Therefore, we explore several machine learning algorithms which have been successfully used for many other NLP tasks and compare the models to find a best model for customer satisfaction classification. Specifically, we apply the following four classification methods: Decision Tree, Naive Bayes, Logistic Regression (a.k.a., maximum entropy classifier), and Support Vector Machines (SVMs).

Decision Tree: A decision tree is a predictive model, which creates a tree providing a mapping from observations about an item (i.e., attributes) to its target value (i.e.. class). In this work, we use C4.5 which builds decision trees using the concept of information entropy [26].

Naive Bayes: Naive Bayes algorithm is a simple probabilistic classifier based on applying Bayes' theorem. The method assumes that all features are mutually independent, and parameter estimation for the naive Bayes models uses the method of maximum likelihood [20]. Given features x_i's and the class variable y, naive Bayes assigns a test example $x = (x_1, ..., x_k)$ to the class y with the highest $P(y|x_1, ..., x_k) = P(y) \prod P(x_i|y)$.

Logistic Regression: Logistic Regression models predict the probability of an event (i.e., class) by fitting data to a logistic curve (a.k.a sigmoid curve). Logistic function is described as $f(z) = \frac{1}{1+e^{-z}}$ where $z = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$, and β_i's are regression coefficients. The success of a logistic regression method is dependent on the appropriateness of "sigmoid" to match the known distribution.

SVMs: The main idea of SVMs is to find a hyperplane which splits the positive examples from negative examples with the largest distance in between the two example sets [35]. In this work, we use C-support vector classification (C-SVC) with a radial basis function (RBF) kernel.

4.2 Features

Customer satisfaction survey results typically include verbatim comments in which customers provide detailed explanations on why they are satisfied or dissatisfied (i.e., voice of the customer). Some sample customer comments which contain the reasons of "completely satisfied" and "less than satisfied" are listed in Table 1.

To learn what factors influence customer satisfaction with a contact center service, we analyzed the verbatim comments in 16,500 C-SAT survey results of the automotive company to identify potential features for C-SAT prediction that can be used to build a C-SAT model. Table 2 lists the most frequently mentioned reasons for being completely satisfied. Interestingly, the reasons for being less than satisfied are essentially the opposites of the reasons for being completely satisfied.

We also analyzed sample call transcripts to identify good determinants of customer satisfaction. Example aspects we investigated include the call duration, the number of onholds during a call, sentiment words, competitor mentions,

Why are you less than satisfied with CAC?	Why are you completely satisfied with CAC?
She had no clue what I was asking about. I was asking her about the operations of the navigation system. She had no answers and she put me in touch with the people who sell discs. Even when she transferred me to the guy who sells the discs, he said 'why would they transfer you to me, this is if you want to buy a disc'.	They were very helpful. Very instructive in explaining how to handle the problem I had. Very, very friendly and explaining it to me. No frustration. They were very friendly. Very knowledgable of what the problem was. They directly guided me through the problem I had.
I didn't get what I was told I was going to receive. I'm not satisfied with what went on with my car. I wouldn't have bought the car had I realized.	Well I called and they responded. Although they said within an hour they came in ten minutes instead, it made me very happy.

Table 1: Examples of customers' feedback provided during a customer satisfaction survey. CAC stands for Customer Assistance Center.

Reasons	Example Quotes
Problem Resolution	tried their best to help, took care of my needs, problem resolved, gave me an accurate answer
Polite CSR	very courteous/polite, was concerned with my problems, listened well, treated me real nice
Knowledgeable CSR	very informative, provided all the information, knowledgeable about vehicle
Responsible CSR	gave her own extension for further questions, made call back, did a follow-up
Speed	handled my issue in a prompt and fast way, gave me a quick answer/resolution
Compensation	got a goodwill, they repaired with no charge

Table 2: Top reasons for being "completely satisfied with the customer assistance center" quoted by the C-SAT survey participants. CSR stands for customer service representative.

and talk speed of the speakers, etc. Based on the analysis of the verbatim comments and the call transcripts, we select the following 20 features which show high correlation with C-SAT scores. The features are categorized into structured, prosodic, lexical and contextual features based on the knowledge sources.

4.2.1 Structured features

Structured features include features that are not usually available in call transcripts, but can be extracted from the contact center's database. Our analysis show that the following two structured features are highly correlated with C-SAT scores.

Goodwill: This feature provides information on whether a goodwill token was offered to the customer, and the type of goodwill offered.

Previous Inbound Interactions: Inbound interactions include any customer-initiated contacts to the contact center. Examples of inbound interactions are calls, emails or instant messaging which the customer initiated. This feature is the number of previous inbound interactions the customer has made before the telephone conversation.

4.2.2 Prosodic features

Prosodic attributes of a conversation provide valuable information about the nature of call, and have widely been used in speech act and dialogue understanding [8, 33]. These attributes can imply the emotional status of the speakers. In this work, we extract the following six prosodic features which can indicate a customer's satisfaction level, and are available in call transcripts. Please note that the presented system is not integrated with an ASR system, and, thus, prosodic features that can only be extracted from acoustic signals such as energy, pitch and F0 are not used.

Long Pause: Long pauses during a call can influence the flow of conversation. For instance, many long pauses by the agent can annoy the customer. In this work, we define a long pause as a pause between two adjacent words lasting more than 5 seconds. The number of all long pauses during a call is used as a feature for classification.

Call Dominance: This feature represents who dominated the conversation in terms of the talking time. Our study found that dissatisfied customers tend to dominate the calls more than satisfied customers.

The call dominance rate is computed based on the relative talking time between the speakers. The talking time of each speaker ($TalkingTime(S_i)$) during a call is computed using the following equation.

$$TalkingTime(S_i) = \sum_{j=1}^{n} TimeDuration(U_{ij})$$

where U_{ij} denotes the j-th utterance spoken by speaker S_i.

The call dominance rate of a speaker S_i, $D(S_i)$, is computed as the percentage of the speaker's talking time over the talking time of all speakers.

$$D(S_i) = \frac{TalkingTime(S_i)}{\sum_k TalkingTime(S_k)}$$

In this work, we use the call dominance rate of the customer as a feature.

Talking Speed: This feature measures the average talking speed of a speaker. The average talking speed of a

speaker is computed by the number of words spoken by the speaker divided by the speaker's talking time in the call.

Our analysis on the speakers' average talking speed reveals interesting insights. Agents tend to talk faster in calls that were reported to be "satisfied" calls than in calls reported to be "dissatisfied" calls (average speed 1.9 in "satisfied" calls vs. 1.5 in "dissatisfied" calls). On the other hand, customers tend to speak faster during "dissatisfied" calls (2.5 in "satisfied" calls vs. 2.8 in "dissatisfied" calls). In this work, the talking speed of both the customer and the CSR are included in the feature set.

Barge-in: Interrupting during the other person's speech may indicate that the person is loosing patience. When an utterance starts before the previous utterance ends, we regard the utterance as a "barge-in". The numbers of barge-ins initiated by both the CSR and the customer are included in the feature set.

4.2.3 Lexical features

Previous work on spoken dialogue analysis mostly include word n-grams as lexical features [8, 32]. In this work, lexical features consist of words which may indicate the customer's emotional state and class-specific words which can reliably distinguish one class from the others. We extract the following eight lexical features.

Product Name: This feature specifies the product family name (in this work, the make of the vehicle) for which the customer is seeking a solution. Typically, customers reveal the product name when they describe the problem they are experiencing.

In this work, we apply a heuristic method using a product taxonomy to identify the product name in call transcripts. We select the first product name mention in the customer's utterances as the product of interest. If no product name is found in the customer's utterances, the first product name mentioned by the CSR is selected. In the case of the automotive company, customers often mention the vehicle's model name, but not the make. We infer the make name using a product taxonomy that provides the relationships between the models and the makes. When no product name is present in the call transcript, the product with most customers is used as the default value.

Filler: Fillers are words or sounds that people often say unconsciously that add no meaning to the communication. Examples of fillers in English include "ah", "uh", "umm", etc. The frequency of fillers in a conversation is often reflective of a speaker's emotional state. Most contact centers encourage their CSRs to minimize the use of fillers. In this work, the numbers of fillers spoken by the customer and the CSR are counted separately, and both numbers are used as features.

Competitor Name: Mentions of competitors or a competitor's product are a good indicator of the customer dissatisfaction with the product. For instance, an unhappy customer might say "I will buy a XXX [1] next time". This sentence does not contain any explicit sentiment, but it certainly expresses a negative sentiment. In this work, we use a manually compiled lexicon of all automotive companies and their product names to recognize competitor mentions. Only the number of competitors' names mentioned by the customer is used.

[1] a competitor's name

Sentiment Word: Call center conversations also contain many words showing the speaker's emotion or affect. To identify words with sentiment polarity, we use the subjectivity lexicon described in [36]. The lexicon contains a list of words with a priori prior polarity (*positive*, *negative*, *neutral* and *both*) and the strength of the polarity (*strongsubj* vs. *weaksubj*). In this work, we use only words of which prior polarity is either *positive* or *negative*, and the strength of the polarity is *strongsubj*. A few words which are frequently used non-subjectively in conversational text such as "okay", "kind", "right", and "yes" are removed from the sentiment word list.

We perform a local context analysis to decide the polarity of a sentiment word (see [36] for more complete contextual polarity analysis) in a context. If a sentiment word has a polarity shifter within a two word window in the left, the polarity of the word is changed based on the shifter [25]. For instance, if a positive sentiment word appears with a negation word, the polarity of word in the context is negative. The number of positive sentiment words and the number of negative sentiment words spoken by the customer are included in the feature set.

Category-specific Word: Some set of words tend to appear more frequently in a certain category than other categories and, thus, can reliably identify the category. We call these words category-specific words. Category-specific words are automatically extracted based on Shannon's entropy, which is a measure of the degree of randomness or uncertainty [30]. More specifically, we define category-specific words as words that appear frequently in the corpus and have low entropy.

The entropy of a word is computed as follows. We first created a corpus of call transcripts, which comprises only the last calls of service requests with manual customer satisfaction survey results. [2] We then calculate the probability of a word, w, appearing in the "satisfied" category (i.e., C-SAT score '4' or '5') and the probability of w appearing in the "dissatisfied" category (i.e., C-SAT score '1' or '2').

$$p_s(w) = \frac{f_s(w)}{f(w)}, \quad p_d(w) = \frac{f_d(w)}{f(w)}$$

where $f_s(w)$ and $f_d(w)$ denote the counts of word w in the "satisfied" call set and in the "dissatisfied" call set respectively, and $f(w) = f_s(w) + f_d(w)$.

The entropy of w, $H(w)$, is defined as in Equation 1.

$$H(w) = -\sum_{i=\{s,d\}} p_i(w) \cdot log_2 p_i(w) \quad (1)$$

In this work, we select words that appear 20 times or more in the corpus, and the entropy is equal to or less than 0.9 (i.e., words appearing in a category 68% or more of the time) as category-specific. Furthermore, if $p_s(w)$ is bigger than $p_d(w)$, the word w is regarded as a "satisfied" word , and otherwise as a "dissatisfied" word. The numbers of "satisfied" words and "dissatisfied" words spoken by the customer are used as features.

[2] we hypothesize that customer satisfaction is more influenced by the last call than earlier calls

Satisfied words	model, ignition, pressure, *BRAND-S*, *MODEL-M*, press, field, mission, *RAS*, cap, update, key, registration, reset, roadside, pennsylvania, glad, *MODEL-S*, reference, reimbursement, park, page, goal, district, attach, march, level, accord, sensor, navigation
Dissatisfied words	lawyer, assembly, lemon, test, damage, paint, conditioner, woman, highway, engineer, sunday, email, court, rid, slow, panel, windshield, crack, indicate, plug, die, safety, store, supervisor, law, responsibility, quality, dollar, report, factory

Table 3: The 30 most category-specific words for each category. We anonymized the automobile brand and model names which can reveal the identify of the company; *BRAND-S* is a luxury brand of the company, *MODEL-M* and *MODEL-S* are two high-end car models, and *RAS* denotes the company's roadside assistance service. Note that more brand names and technical words appear in the "satisfied" category, and "dissatisfied" category contains more legal terms such as "lawyer" and "court".

Table 3 lists the most "satisfied" words and "dissatisfied" words.

4.2.4 Contextual features

Contextual features are phrases or expressions used in certain contexts which can affect the customer's satisfaction level. Based on our analysis of customers'comments and sample call transcripts, we identified the following four contextual features.

CSR's Positive Attitude: These features intend to reflect the CSR's positive attitude toward the customer. We manually collected a list of phrases which CSRs often use to express courteousness or to rephrase the customer's problem. For instance, "let me see if I understood..." and "as I understand, ..." can hint that the CSR is trying to understand the customer's question correctly. Also, expressions like "I am happy to assist/resolve/address .." and "I am sorry to hear .." in the beginning of a call can indicate that the CSR was sympathetic and willing to help the customer. In this work, we count the number of such expressions in the first ten utterances spoken by the CSR.

CSR's Contact Information: As noted in Table 2, customers consider a CSR responsible when the CSR provided her contact information for the customer to be able to reach the CSR directly in a later time. Example of the expressions are "further question", "my number", "contact information", "extension", and "call me back". We recognize these expressions in the last ten utterances spoken by the CSR.

Follow-up Schedule: A follow-up is a call made by the CSR to the customer after the current call is ended. We can not know from the transcript of the current call if there was a follow-up. Instead, we check if the CSR scheduled a follow-up during the conversation.

A follow-up schedule can be an attribute for a responsible CSR, but also can indicate that the customer's problem was not resolved during the call. CSRs usually schedule a follow-up at the end of the call, and obtain the customer's contact information. We recognize the existence of a follow-up schedule by identifying cue words such as "call you back" and "touch base" and expressions for a telephone number, day and hour information in the last 20 utterances.

Gratitude: Finally, we look at the customer's response at the end of the call. When the customer uses many expressions showing gratitude such as "appreciate" and "great", that can indicate that the customer is satisfied. We count the number of such expressions in the last ten utterances spoken by the customer.

5. EXPERIMENTS

As described in Section 3, we conduct experiments for 5-point satisfaction classification and 2-point satisfaction classification using two sets of customer calls; a call transcript set comprising entire conversations and a call transcript set comprising only the first half of conversations. Especially, the fist half conversations are used to investigate the feasibility of measuring customer satisfaction in real-time when the conversation is still in progress.

The experiments were conducted with RapidMiner, a machine learning toolkit offering a wide range of methods for data pre-processing, machine learning and validation [19]. We used the default settings in RapidMiner for Decision Tree and Naive Bayes. For Support Vector Machines, we use the C-support vector classification (C-SVC) with a radial basis function (RBF) kernel as implemented in LIBSVM [5]. Logistic Regression uses maximum likelihood, which is an iterative procedure. We set the maximum number of iterations to 300. The standard Logistic Regression in RapidMiner applies only to binary classification, and it was extended to a multiclass classifier for the 5-point satisfaction classification using "one-against-all" strategy [29].

5.1 Data

We acquired customer calls to a contact center of the automotive company which were recorded during a two month period time in 2007. The call set constitutes the base source of our experimental data. The calls were transcribed using the IBM Attila Speech recognition toolkit [31]. The ASR system was retrained with sample customer calls from the same contact center as well as general conversational telephony speech data and broadcast news, and shows an overall word error rate of 26%.

To develop supervised machine learning systems, we need annotated ground truth data. Hand annotation of customer satisfaction is not only time consuming but also very difficult. Customer satisfaction is very subjective and, thus, is hard to achieve high inter-annotator agreement as experienced in previous work on sentiment analysis [8, 32, 6]. To avoid the need of costly and inconsistent human annotation, we use manual C-SAT survey results as the ground truth. We argue that the satisfaction ratings in the surveys are in fact hand annotation done by the customer themselves and, thus, most accurate.

We obtained the manual C-SAT survey results conducted for the calls used in this work by matching the surveys with

the customer calls. Note that a C-SAT survey is conducted for a service request not for an individual call. A service request typically consists of multiple interactions between a customer and one or more agents via multi-modal media including telephone conversations, emails and postal mail. In many cases, a service request comprises more than one telephone conversations resulting in a 1-to-n relationship between a C-SAT score and customer calls. A C-SAT score for a service request reflects the customer's cumulative experience across multiple interactions with the contact center.

To mitigate this problem, we selected the service requests which involved only one incoming call from the respective customers, resulting in 115 service requests. The cumulative call length of the 115 calls is 27 hours 34 minutes 55 seconds, and the call transcripts contain 171,860 tokens and 16,323 speaker turns. Among 16,323 utterances, 8,139 utterances were spoken by the CSRs, and 8,184 utterances were spoken by customers showing almost same talk distribution by the CSRs and the customers.

5.2 Baseline Systems

In this work, we use the following two baseline systems for performance evaluation purpose. The first baseline system is an artificial classifier which assigns all calls to the most frequent class (i.e., C-SAT score "5" for 5-point classification, and "satisfied" for 2-point classification). This baseline is called *DominantClass* hereafter. Table 4 shows the distribution of the 115 calls across the five C-SAT scores and the three categories.

C-SAT Score	Category	Number of Calls
1	*Dissatisfied*	19
2		6
3	*Neutral*	11
4	*Satisfied*	13
5		66

Table 4: The number of calls across numerical C-SAT scores and three categories. The accuracy of the *DominantClass* baseline is 57.39% and 75.96% for 5-point classification and 2-point classification respectively.

The second baseline comes from the CSRs who handled the customer calls. In the contact center, the CRSs are required to judge if the customer is satisfied or dissatisfied when they close a service request, and to record their judgment in the database. We use the CSRs' judgment as the second baseline system, and call it *CSRJudgment*. Note that the CSRJudgment baseline can only be used for the 2-point satisfaction classification.

Table 5 shows the contingency table of the CSRs' judgment on customer satisfaction. As we can see in the table, CSRs identified satisfied customers with high precision and recall, but recall for dissatisfied customers is very low.

5.3 Measuring Customer Satisfaction at the End of Calls

In this section, we discuss the experimental results of the 5-point satisfaction classification and the 2-point satisfaction classification at the end of calls. The performance of the automatic systems are compared with the two baseline systems based on average classification accuracy, precision, recall and F_1-measure of 10-fold cross validation.

		True		
		Satisfied	Dissatisfied	Precision
CSR	Satisfied	74	9	89.16%
	Dissatisfied	5	16	76.19%
Recall		93.67%	64%	

Table 5: The contingency table of the customer service representatives' judgment on customer satisfaction. The accuracy of the *CSRJudgment* baseline is 86.54%, and F_1 measures for "satisfied" calls and "dissatisfied"calls are 91.36% and 69.57% respectively.

5.3.1 Performance of 5-point satisfaction classification

The accuracy of the baseline system and the four classification systems for 5-point C-SAT classification are summarized in Table 6. The second column (**All**) displays the best performance of each algorithm when all features were used. The accuracy reported here is the average accuracy of 10-fold cross validation.

Methods	Classification Accuracy
DominantClass	57.39
Decision Tree	60.87
Logistic Regression	59.13
Naive Bayes	60.00
SVM	66.09

Table 6: Accuracy for 5-point C-SAT classification. All numbers are in percentage.

As we can see from the table, all four automatic methods outperform the *DominantClass* baseline. The SVM-based approach achieves the best accuracy (66.09%) which outperforms the baseline method by over 15%. There is no substantial performance difference among the other three approaches.

5.3.2 Performance of 2-point satisfaction classification

The experimental results of 2-point satisfaction classification are described in Table 7 in detail. For 2-point satisfaction classification, we compare the four automatic systems with both the *DominantClass* baseline and the *CSRJudgment* baseline in terms of classification accuracy, precision, recall and F-measure.

The highest classification accuracy for 2-point satisfaction classification (89.42%) was achieved by the decision tree-based approach and the SVM-based approach. The methods outperform the *DominantClass* baseline and the *CSRJudgment* baseline by 17.7% and 3.3% respectively. Furthermore, both systems produce higher F-measure values than the *CSRJudgment* baseline in identifying both satisfied calls and dissatisfied calls. Specially note that the decision tree-based system achieves 11.5% higher F-measure than the human judgment for identifying dissatisfied calls.

5.4 Effect of Features

In this section, we discuss the relative contributions of the different feature types to automatic C-SAT measurement. We ran the experiments with one feature type removed at a

Methods	Classification Accuracy	*Satisfied* Calls			*Dissatisfied* Calls		
		Precision	Recall	F-measure	Precision	Recall	F-measure
DominantClass	75.96	75.96	100.0	86.34	0.00	n/a	n/a
CSRJudgment	86.54	89.16	93.67	91.36	76.19	64.00	69.57
Decision Tree	89.42	92.50	93.67	93.08	79.17	76.00	77.55
Logistic Regression	85.58	92.41	97.47	90.68	68.09	64.00	68.09
Naive Bayes	83.65	82.98	98.73	90.17	90.00	36.00	51.43
SVM	89.42	87.78	100.0	93.49	100.0	56.00	71.79

Table 7: Comparison of precision, recall and F_1 measure of the two baseline systems and the four automatic systems for 2-point C-SAT classification. All the numbers are in percentage.

time (i.e., leave-one-out), and compare the results. **All** indicates that all features were used. **All-Str**, **All-Pro**, **All-Lex**and **All-Con** indicate the cases where the structured, prosodic, lexical and contextual features were removed respectively.

The comparison of classification accuracy for 5-point and 2-point classification with the different feature sets are depicted in Figure 2 and Figure 3 respectively. As we can see from the charts, the **All** model outperforms all other models except Decision Tree's **All-Pro** model for 2-point satisfaction. Also note that structured and lexical features have bigger impact on C-SAT measurement than the other two feature types.

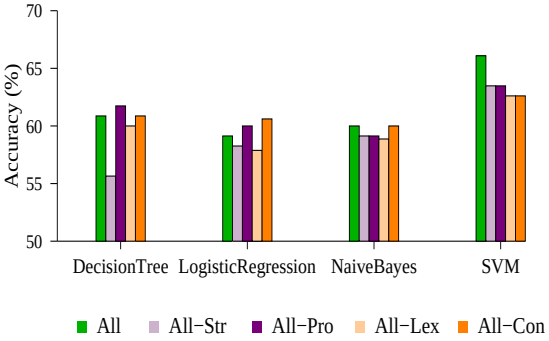

Figure 2: Effects of the different feature sets on 5-point satisfaction classification

5.5 Measuring Customer Satisfaction in the Middle of Calls

In the previous section, we showed that customer satisfaction can be automatically measured with high accuracy by analyzing by analyzing the conversation between a customer and a CSR. Another interesting question is that if we can "predict" C-SAT in real-time, i.e., when the conversation is still in progress. With a real time transcription system, such tools can enable supervisors take over a call when a customer becomes unhappy to resolve the customer's issue.

To answer this question, we conduct experiments with only the first half of the conversations and measure how accurately we can predict C-SAT in the middle of a call. Since SVMs and Decision Tree methods were proven to be best performing approaches, we carried out the experiments only with the two approaches. Furthermore, it is worth noting that "Goodwill" information is typically available at the

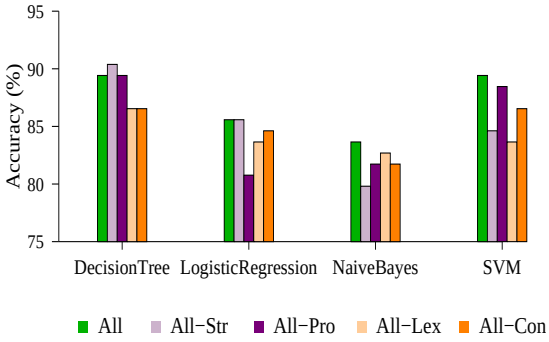

Figure 3: Effects of the different feature sets on 2-point satisfaction classification

end of calls, and thus we removed "Goodwill" feature from the feature set. All other feature values were extracted from the automatically transcribed transcripts of the first half of the calls.

Figure 4 depicts the accuracy comparison of customer satisfaction measurement using the first half of calls with the results obtained from using the entire calls, for both 5-point satisfaction classification and 2-point satisfaction classification. Both Decision Tree and SVM perform better when the entire conversations were available. However, both methods significantly outperform the *DominantClass* baseline even with only the half of calls. Also note that the SVM method produces the accuracy comparable to the *CSRJudgment* baseline for the 2-point satisfaction classification.

Figure 5 shows the comparison of F-measure of 2-point satisfaction classification. The results confirm that analyzing the entire conversations provides more accurate prediction of customer satisfaction than analyzing only partial conversations. As we can see from the figure, the degree of performance degradation is larger for "dissatisfied" calls than for "satisfied" calls. Also, SVMs are shown to be less prone to the information loss than Decision Tree for identifying both "satisfied" calls and "dissatisfied" calls.

The main reasons of the performance degradation might be the following. First, the "Goodwill" feature is not used at all for the experiment with the first half of calls. Second, contextual features including "CSR's Contact Information", "Follow-up Schedule" and "Gratitude" typically appear at the end of calls. Therefore, these features are mostly absent for C-SAT prediction in real-time. It is worth noting that the presence of "CSR's Contact Information" and "Follow-up

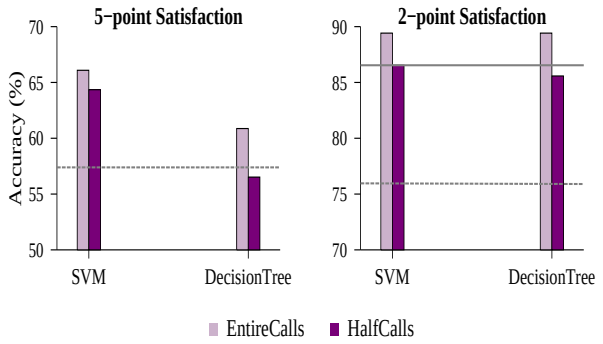

Figure 4: Accuracy comparison of customer satisfaction measurement with two different sets of customer calls. "EntireCalls" denotes the classification results from using the entire conversations, and "HalfCalls" shows the classification results from analyzing only the first half of the conversations. The dashed line denotes the accuracy of the *DominantClass* baseline, and the solid line denotes the accuracy of the *CSRJudgment* baseline.

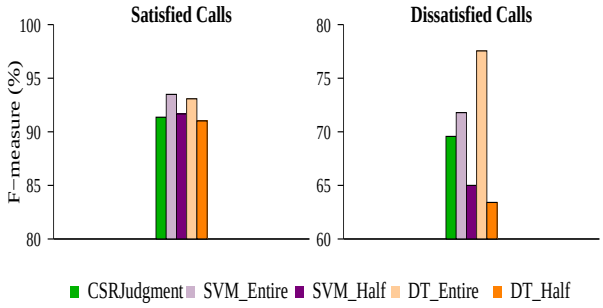

Figure 5: Comparison of F-measure for "Satisfied" calls and "Dissatisfied" calls in the cases where the entire calls were used and where only half of the calls were used. 'DT' stands for 'DecisionTree' in the legend. '_Entire' and '_Half' denotes the cases where the entire conversation set and where the half conversations were used respectively.

Schedule" often indicate that the customer's issue was not resolved during the call. The results seem to support the findings by other research on the strong correlation of first call resolution and customer satisfaction [7, 21].

6. CONCLUSIONS AND FUTURE WORK

Customer satisfaction is one of the key performance indicators of contact centers. However, due to high cost, contact centers conduct a manual survey with a very small number of customers limiting the value of the survey results. The primary goal of this work is to investigate if customer satisfaction can be automatically measured by analyzing automatically generated call transcripts using NLP and machine learning (ML) technologies. Such tools can enable companies to measure customer satisfaction for each and every call in near real-time, and, thus, to obtain more reliable knowledge about customer satisfaction.

We analyzed manual customer satisfaction survey results and sample call transcripts to identify features that are highly correlated with customer satisfaction scores. Analysis of such survey results can provide features for C-SAT prediction that can be used to build a C-SAT model for predicting C-SAT with high accuracy. Our experiments show that automatic C-SAT measurement using machine-generated call transcripts is feasible. Automatic C-SAT measurement at the end of calls outperform human judgment in terms of both overall classification accuracy and F-measure. Experiments for measuring customer satisfaction in real-time, i.e., while the conversation is still in progress, also produce classification accuracy comparable to human judgment with much less information. The results imply that, with a real time transcription system, such tools can allow supervisors take over a call when a customer becomes unhappy and to resolve the customer's issue directly preventing customer defection.

To further improve the accuracy of automatic C-SAT measurement, we plan to extend the feature set to include acoustic features such as F0, pitch and energy level of the voices, call information such as the call waiting time, and call history information such as if a promised follow-up call was actually made.

7. REFERENCES

[1] Anderson, E. W., C. Fornell, and R. T. Rust: 1997, 'Customer Satisfaction, Productivity, and Profitability: Differences Between Goods and Services'. *MARKETING SCIENCE* **16**(2), 129–145.

[2] Brennan, M., S. Benson, and Z. Kearns: 2005, 'The effect of introductions on telephone survey participation rates'. *International Journal of Market Research* **47**(1), 65–74.

[3] Burton, S. M.: 1997, 'Modelling the Determinants of Customer Satisfaction'. In: *Ph.D. thesis, The University of New South Wales, Australia.*

[4] Byrd, R. J. N. M. S., W. Teiken, Y. Park, K. F. Cheng, S. C. Gates, and K. Visweswariah: 2008, 'Semi-automated logging of contact center telephone calls'. In: *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management.* pp. 133–142.

[5] Chang, C. and C. Lin: 2001, 'LIBSVM: a library for support vector machines'. Software available at *http://www.csie.ntu.edu.tw/ cjlin/libsvm.*

[6] Devitt, A. and K. Ahmad: 2008, 'Sentiment Polarity Identification in Financial News: A Cohesion-based Approach'. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL).* pp. 984–991.

[7] Feinberg, R., I.-S. Kim, L. Hokama, K. de Ruyter, and C. Keen: 2000, 'Operational determinants of caller satisfaction in the call center'. *International Journal of Service Industry Management* **11**(2), 131–141.

[8] Forbes-Riley, K. and D. Litman: 2004, 'Predicting Emotion in Spoken Dialogue from Multiple Knowledge Sources'. In: *HLT-NAACL 2004: Main Proceedings.* pp. 201–208.

[9] Gamon, M.: 2004, 'Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis'. In: *Proceedings of Coling 2004*. pp. 841–847.

[10] Godbole, S. and S. Roy: 2008, 'Text Classification, Business Intelligence, and Interactivity: Automating C-Sat Analysis for Services Industry'. In: *Proceeding of the 14th ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD)*. pp. 911–919.

[11] Hallowell, R.: 1996, 'The relationships of customer satisfaction, customer loyalty, and profitability: an empirical study'. *International Journal of Service Industry Management* **7**, 27–42.

[12] Hu, M. and B. Liu: 2004, 'Mining and Summarizing Customer Reviews'. In: *Proceedings of the Tenth International Conference on Knowledge Discovery and Data Mining*. pp. 168–177.

[13] Kanayama, H. and T. Nasukawa: 2006, 'Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 355–363.

[14] Kim, S.-M. and E. Hovy: 2006a, 'Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text'. In: *Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text*. pp. 1–8.

[15] Kim, S.-M. and E. Hovy: 2006b, 'Identifying and Analyzing Judgment Opinions'. In: *Proceedings of the Human Language Technology Conference of the NAACL*. pp. 200–207.

[16] Kobayashi, N., K. Inui, and Y. Matsumoto: 2007, 'Extracting Aspect-Evaluation and Aspect-Of Relations in Opinion Mining'. In: *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. pp. 1065–1074.

[17] Likert, R.: 1932, 'A Technique for the Measurement of Attitudes'. *Archives of Psychology* **140**. pp 1–55.

[18] Liscombe, J., J. Venditti, and J. Hirschberg: 2003, 'Classifying subject ratings of emotional speech using acoustic feature'. In: *Proceedings of EUROSPEECH*. pp. 725–728.

[19] Mierswa, I., M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler: 2006, 'YALE: Rapid Prototyping for Complex Data Mining Tasks'. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 935–940.

[20] Mitchell, T.: 1997, *Machine Learning*. McGraw Hill.

[21] Monger, J., M. Rudick, and L. O'Flahavan: 2004, 'First call resolution: its impact and measurement'. *Contact Professional*. pp. 24–27.

[22] Nambiar, U., H. Gupta, and M. Mohania: 2007, 'CallAssit: Helping Call Center Agents in Preference Elicitation'. In: *Proceedings of VLDB*. pp 1338–1341.

[23] Pang, B. and L. Lee: 2005, 'Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales'. Ann Arbor, Michigan, pp. 115–124.

[24] Pang, B., L. Lee, and S. Vaithyanathan: 2002, 'Thumbs up? Sentiment Classification using Machine Learning Techniques'. In: *Conf. on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 79–86.

[25] Polanyi, L. and A. Zaenen: 2004, 'Contextual Valence Shifters'. In: *Working Notes – Exploring Attitude and Affect in Text: Theories and Applications*.

[26] Quinlan, J. R.: 1993, *Programs for Machine Learning*. Morgan Kaufmann Publishers.

[27] Ranaweera, C. and J. Prabhu: 2003, 'The influence of satisfaction, trust and switching barriers on customer retention in a continuous purchasing setting'. *International Journal of Service Industry Management* **14**. pp. 374–395.

[28] Reichheld, F. F.: 2001, *Loyalty Rules: How Today's Leaders Build Lasting Relationships*. Harvard Business School Press.

[29] Rifkin, R. M. and A. Klautau: 1998, 'In defense of one-vs-all classification'. *Journal of Machine Learning* **5**. pp 101–141.

[30] Shannon, C. E.: 1948, 'A Mathematical Theory of Communication'.

[31] Soltau, H., B. Kingsbury, L. Mangu, D. Povey, G. Saon, and G. Zweig: 2004, 'The IBM 2004 conversational telephony system for rich transcription'. In: *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*. pp. 205–208.

[32] Somasundaran, S., J. Ruppenhofer, and J. Wiebe: 2007, 'Detecting Arguing and Sentiment in Meetings'. In: *SIGdial Workshop on Discourse and Dialogue*.

[33] Takeuchi, H., L. Subramaniam, T. Nasukawa, , and S. Roy: 2007, 'Combining Multiple Knowledge Sources for Dialogue Segmentation in Multimedia Archives'. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. pp. 1016–1023.

[34] Turney, P. D.: 2002, 'Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews'. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. pp. 417–424.

[35] Vapnik, V.: 1995, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

[36] Wilson, T., J. Wiebe, and P. Hoffmann: 2005b, 'Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. pp. 347–354.

[37] Yi, J., T. Nasukawa, R. Bunescu, and W. Niblack: 2003, 'Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques'. In: *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM-2003)*. pp. 427–434.

[38] Zweig, G., O. Siohan, B. Ramabhadran, D. Povey, L. Mangu, and B. Kingsbury: 2006, 'Automated Quality Monitoring in the Call Center with ASR and Maximum Entropy'. In: *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*.