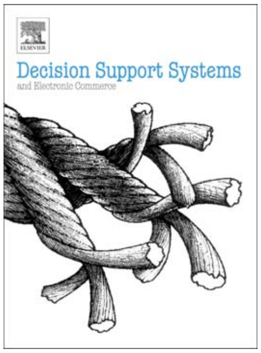
Accepted Manuscript



Modeling Customer Satisfaction from Unstructured Data Using A Bayesian Approach

Mohsen Farhadloo, Raymond A. Patterson, Erik Rolland

PII: S0167-9236(16)30102-6
DOI: doi: 10.1016/j.dss.2016.06.010
Reference: DECSUP 12734

To appear in: *Decision Support Systems*

Received date: 24 April 2015
Revised date: 3 June 2016
Accepted date: 14 June 2016

Please cite this article as: Mohsen Farhadloo, Raymond A. Patterson, Erik Rolland, Modeling Customer Satisfaction from Unstructured Data Using A Bayesian Approach, *Decision Support Systems* (2016), doi: 10.1016/j.dss.2016.06.010

Modeling Customer Satisfaction from Unstructured Data Using A Bayesian Approach

Mohsen Farhadloo

School of Engineering, University of California, Merced

Raymond A. Patterson

Haskayne School of Business, University of Calgary, Canada

Erik Rolland

E&J Gallo Management Program, School of Engineering, University of California, Merced

mfarhadloo@ucmerced.edu, raymond.patterson@ucalgary.ca, erolland@ucmerced.edu

Abstract

The Internet is host to many sites that collect vast amounts of opinions about products and services. These opinions are expressed in written language, and this paper presents a method for modeling the aspects of overall customer satisfaction from free-form written opinions. Written opinions constitute unstructured input data, which are first transformed into semi-structured data using an existing method for aspect-level sentiment analysis. Next, the overall customer satisfaction is modeled using a Bayesian approach based on the individual aspect rating of each review. This probabilistic method enables the discovery of the relative importance of each aspect for every unique product or service. Empirical experiments on a data set of online reviews of California State Parks, obtained from TripAdvisor, show the effectiveness of the proposed framework as applied to the aspect-level sentiment analysis and modeling of customer satisfaction. The accuracy in terms of finding the significant aspects is 88.3%. The average R^2 values for predicted overall customer satisfaction using the model range from 0.892 to 0.999.

Keywords: Aspect-level sentiment analysis, Customer satisfaction modeling, Bayesian framework.

Preprint submitted to *June 22, 2016*

1. Introduction

Organizational decision-making increasingly relies on Decision Support System (DSS) tools. During the past 30 years, research on machine learning has enabled these DSS tools to become progressively more intelligent [1]. In particular, machine learning algorithms have enabled DSS to learn and to be responsive to changing decision-making environments. In this paper, we aim to expand on how machine learning can be applied to a dynamic decision environment where we can improve our understanding of customer satisfaction based on online product or service reviews.

Customer review websites, such as TripAdvisor for travel, Yelp and Urban Spoon for restaurants, and Patagonia, Lands' End and Epinions for product reviews have become commonplace. These sites allow customers to both read and provide reviews of a product or service. The customer reviews, which are often openly available on the web, contain a wealth of information usable by management, competitors, investors, and other stakeholders to discover the customer concerns which drive overall customer satisfaction for a particular service or product. For simplicity, we use the terms ”overall satisfaction” and ”overall rating” interchangeably. The term ”aspect” denotes a real or perceived component or feature of a product or a service that matters to the customer. Traditionally, customer satisfaction questionnaires would be used to determine the significant components, or aspects, of overall customer satisfaction [2]. However, questionnaires are expensive or may not be available. In some cases, public agencies are even prohibited by law from collecting satisfaction questionnaires from customers. In situations such as this, the only alternative may be to analyze publicly available free-form text comments, for example, from sources such as TripAdvisor. Even in cases where customer satisfaction surveys are collected, analysis of free-form text comments available on the Internet can provide important insights usable in isolation or in conjunction with traditional closed-form customer satisfaction questionnaires.

An ideal algorithm is one which would allow for aspect identification without *a priori* knowledge regarding the aspects, and which then would determine the significance of the aspects' impact on overall customer satisfaction. The benefits of such a system would be that, once trained, the algorithm would be able to discover the components and their relative importance on overall satisfaction without further intervention. Utilizing a prior method for identifying aspects proposed by Farhadloo and Rolland [3], this paper proposes a Bayesian approach for modeling the overall customer satisfaction based on the individual aspect rating of each review.

Two major challenges must be addressed when attempting to relate written text comments to customer satisfaction. First, the written comments must be prepared for analysis. This is done by extracting aspects and sentiments from the written comments. Figure 1 illustrates this data-transformation process, and it will be discussed in further detail in section 4. Second, the importance of the various components and their impact on overall customer satisfaction must be ascertained. This challenge is addressed with the development of a Bayesian method that determines the rating of each aspect, and the relative importance of the aspects, for each individual product or service. This method also allows a prediction to be made regarding the overall customer satisfaction of each contributor for use in situations where comments are provided but the overall satisfaction rating is unavailable. Empirical experiments on a data set of online reviews demonstrate the effectiveness of the proposed methods.

2. Related Work

Customer satisfaction can be viewed as a constellation of aspects and associated sentiments. For example, an aspect for a restaurant service could be the "bathroom". The sentiments associated with the bathroom aspect could drive the satisfaction up or down (e.g., "clean" versus "dirty"). The research challenges associated with written comments would then be to first understand what the aspects are, as well as the associated sentiments about the found aspects.

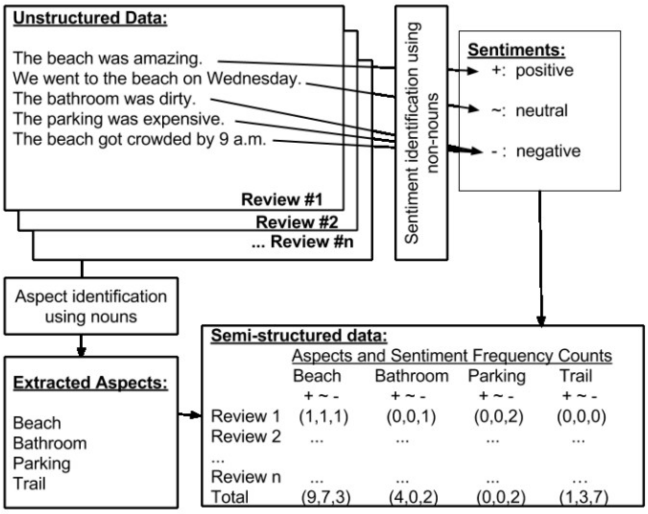

Unstructured Data:
The beach was amazing.
We went to the beach on Wednesday.
The bathroom was dirty.
The parking was expensive.
The beach got crowded by 9 a.m.
Review #1
Review #2
... Review #n
Sentiment identification using non-nouns
Sentiments:
+: positive
~: neutral
- : negative
Aspect identification using nouns
Extracted Aspects:
Beach
Bathroom
Parking
Trail
Semi-structured data:
Aspects and Sentiment Frequency Counts
Beach Bathroom Parking Trail
+ ~ - + ~ - + ~ - + ~ -
Review 1 (1,1,1) (0,0,1) (0,0,2) (0,0,0)
Review 2 ... ...
...
Review n ... ...
Total (9,7,3) (4,0,2) (0,0,2) (1,3,7)

Figure 1: System flowchart for transforming the unstructured data into semi-structured data.

We refer to this as the first challenge. In short, this process takes unstructured written comments and converts them into semi-structured data that can be utilized for further analysis. Next, we then seek to understand the relative contribution of each aspect with respect to customer satisfaction, and refer to this as our second challenge.

Addressing the first challenge of giving structure to unstructured text data, Moghaddam and Ester [4] present an unsupervised method for aspect extraction from unstructured reviews using known aspects. That means the aspects that may impact customer satisfaction are presumed to be known *a priori*. This is useful in cases where the customer satisfaction of services and products are well understood, but obviously restricts usefulness only to areas for which aspects are known with a high degree of certainty, and also only in domains where these aspects are unlikely to change over time.

Extending the research to cases where the aspect are not known *a priori*, Farhadloo and Rolland [3] propose a method to identify the aspects using cluster analysis. Their aspect identification is based on a bag of nouns (BON), which means that it is limited to contexts which share the same nouns and aspects as the training set. Although the method proposed by Farhadloo and Rolland [3] does not explain the variance in overall customer satisfaction with respect to the important aspects, the paper proposes an appropriate framework for aspect-level sentiment analysis.

An alternative to the cluster analysis approach described above is topic modeling. Topic models are probabilistic techniques based on hierarchical Bayesian networks for discovering the main themes existing in a collection of unstructured documents [5]. Latent Dirichlet Allocation (LDA) is the simplest type of topic model, which treats the topic mixture weights as a K-dimensional hidden random variable [5]. Various modifications have been proposed to be able to apply LDA to sentiment analysis [6, 7, 8]. Xianghua et al. [9] use topic modeling to create an unsupervised approach to automatically discover both aspects and sentiments. Gan et al. [10] develop a method to infer interactions between topics. Zhou and Carin [11] use a negative binomial process for topic modeling, and show that LDA is a special case of their method. In contrast, Cambria et al. [12] present concept-level sentiment analysis which augments traditional sentiment analysis to better extract opinion from text (e.g., tasks such as microtext analysis, semantic parsing, subjectivity detection, anaphora resolution, sarcasm detection, topic spotting, aspect extraction, and polarity detection), all to create a better semantic extraction tool. In our view, while these additional techniques are interesting, they simply represent alternative methods of extracting aspects or extracting sentiments. Methods to extract aspects and sentiments are secondary to the main point of this paper, which is to use the aspects and sentiments to explain overall customer satisfaction. For a comprehensive literature review of topic modeling, we refer the reader to Farhadloo and Rolland [13].

Consistent with cluster analysis, note that in topic modeling there is no computation of the relative importance of the aspects with respect to overall customer satisfaction. However, this relative importance is essential to full understanding of customer satisfaction. This paper will utilize cluster analysis rather than the topic modeling methodology due to the relative simplicity of connecting the results from the cluster analysis to the second challenge of finding the relative importance of the aspects that make up customer satisfaction.

Addressing the second challenge, the work by Wang, Lu and Zhai [14] presents a maximum likelihood estimation method for determining the significant components of customer satisfaction in free-form text. In Wang et al. [14], the aspects are presumed to be known *a priori*, but in this paper we are using an approach based on sentence clustering to identify aspects. In their approach, the only piece of data that contains rating information is the observed overall rating of each review, and the aspects' ratings are estimated using the overall rating and the frequency of terms that are related to the corresponding aspect. In the Wang et al. approach [14], the rating of each aspect is estimated based on the frequency of the terms related to that aspect, and a $|V|$-dimensional vector must be learned for each aspect that indicates the word sentiment polarities of that aspect. $|V|$ is the size of their vocabulary list, which is usually a large number. This issue demonstrates itself in the parameter estimation phase as well. During the parameter estimation, in order to avoid large matrix inversion ($|V| \times |V|$ matrix), they used nonlinear optimization methods.

The purpose of our paper is to build on the past research by simultaneously creating: 1) a method for handling aspects that are not known or understood a-priori, and 2) to create a more efficient method of determining the relative importance of the found aspects. We utilize the method of Farhadloo and Rolland [3] as a basis for generating semi-structured data without requiring aspects to be known *a priori*. We then develop a method for determining the aspects' relative impact on customer satisfaction, addressing some of the major limitations of Wang et al. [14]. Particularly, we argue that in addition to the overall rating, one could possibly improve on prior research if the frequencies of the positive, neutral and negative sentiment classes associated with each aspect are to be included in the aspect rating estimation. This also has the advantage of dealing with much smaller matrices in the model (a 3×3 matrix), allowing the development of a closed-form solution for updating the mixing coefficient for each aspect. Hence, this research sets out to address situations where aspects are not known a-priori, and the significance of each aspect is efficiently determined based on aspect sentiments.

3. Problem Definition

In this section, we present the research problem, basic definitions, core concepts, sub-problems and target objectives. We also borrowed some of the definitions from prior literature [15].

Contributor: A Contributor is the person who is expressing their opinion in written language or text. The contributor may also be called an opinion holder or opinion source.

Object: An object is an entity which can be a product, service, person, event, organization, or topic [15]. It may be associated with a set of components and attributes. Sentiment analysis literature refers to these components and attributes as aspects.

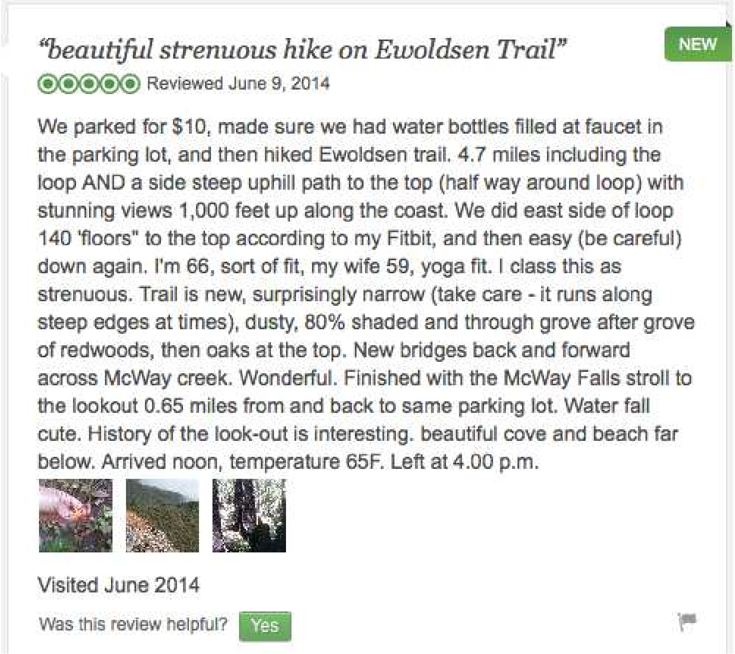
Review: A Review is a contributor-generated text that contains the opinions of the contributor about some aspects of the object.

Opinion: An opinion on an aspect is a positive, neutral or negative view, attitude, emotion or appraisal on that aspect from a contributor.

Aspect: An aspect is an important attribute of the object with respect to overall costumer satisfaction that the contributor has commented on in their review.

In our problem, the input is a set of reviews of some objects, where each review has an overall rating. Such practices are common in customer review websites, and most online review sites contain these components.

Formally, we denote the set of available reviews with $D = \{d_1, d_2, \ldots, d_{|D|}\}$ where each review $d \in D$ comes with an overall rating r_d. Figure 2 shows a typical review by a contributor. In each review, the contributor is expressing his/her opinions about some aspects of the object of interest and also gives an overall rating to the object from their point of view. In this problem we are



NEW

"beautiful strenuous hike on Ewoldsen Trail"

●●●●● Reviewed June 9, 2014

We parked for $10, made sure we had water bottles filled at faucet in the parking lot, and then hiked Ewoldsen trail. 4.7 miles including the loop AND a side steep uphill path to the top (half way around loop) with stunning views 1,000 feet up along the coast. We did east side of loop 140 'floors" to the top according to my Fitbit, and then easy (be careful) down again. I'm 66, sort of fit, my wife 59, yoga fit. I class this as strenuous. Trail is new, surprisingly narrow (take care - it runs along steep edges at times), dusty, 80% shaded and through grove after grove of redwoods, then oaks at the top. New bridges back and forward across McWay creek. Wonderful. Finished with the McWay Falls stroll to the lookout 0.65 miles from and back to same parking lot. Water fall cute. History of the look-out is interesting. beautiful cove and beach far below. Arrived noon, temperature 65F. Left at 4.00 p.m.

Visited June 2014

Was this review helpful? Yes

Figure 2: A sample review from *TripAdvisor.com*

interested in: 1) discovering the aspects that the contributors have commented on based on the content of all reviews in D and 2) discovering the relative importance of those aspects with respect to overall satisfaction for that object from the contributors' perspectives.

4. Methods

A challenge in this research is the transformation of the raw input data, which is available in the form of written reviews, into semi-structured data. Semi-structured data is data that is neither raw data (written reviews), nor data conforming to the formal structure of relational databases or other forms of data tables, but nonetheless contains tags or other markers to separate semantic elements (see Figure 1). To address this challenge, in Appendix A we show how a framework for aspect-level sentiment analysis proposed by Farhadloo and

Rolland [3] can be used to give structure to written review data. The result is the development of semi-structured data as illustrated in Figure 1.

Another major challenge, and the focus of this paper, is to discover the relative importance of the aspects for each object from the contributors' perspectives. Regarding this second challenge, a probabilistic approach is proposed that models the overall satisfaction of each contributor as a combination of the ratings of different aspects for each object. Using this probabilistic model it is possible to identify the relative importance of each aspect, for each object, from the contributors' perspective.

4.1. Probabilistic modeling of overall satisfaction

In this section we provide a probabilistic approach to address the following related questions:

1. How to generate a single rating for each aspect using its positive, neutral and negative sentiment counts,
2. How to discover the relative importance of each aspect from the contributor's perspective.

Note that the transformation of unstructured data into the described semi-structured form (see Appendix A) yields 3 numbers associated with each aspect reflecting the frequency counts for positive, negative, and neutral sentiments. One assumption in our model is that the associated rating of each aspect is based on a particular combination of the positive, neutral and negative sentiments of that aspect. This means that the overall aspect rating depends upon how many times an aspect has been associated with positive, neutral and negative sentiments in a single review. In addition to the plain text of each review, there is an overall rating assigned to each review by the contributor. We assume that the contributor assigns the overall rating as a weighted sum of all of the individual aspect ratings, where the weights in this linear combination reflect the relative importance she or he has put on each aspect. Next, we propose a probabilistic model that captures these assumptions.

For each review d, there is a $3K$-dimensional vector containing the normalized positive, neutral and negative frequency counts associated with each aspect a_{dij}. We denote the overall rating assigned to each review by r_d and assume there are K extracted aspects, then:

$$r_d = \sum_{i=1}^{K}(w_{di} \sum_{j=1}^{3}(\alpha_j a_{dij})) \tag{1}$$

where w_{di} is the weight of the i^{th} aspect in document d, $a_{di1}, a_{di2}, a_{di3}$ are the normalized frequency counts of positive, neutral and negative sentiments of the i^{th} aspect and $\alpha_i, \alpha_2, \alpha_3$ are the (combination) coefficients that combine the aspect's sentimental frequencies in order to get the single aspect rating, respectively. The model in equation (1) can be seen as a regression model in which r_d is the dependent variable and a_{dij} are the independent variables.

In order to take into account any other factor that may derive the overall satisfaction and any uncertainty in the model of equation (1), the overall satisfaction is assumed to have a Gaussian distribution with mean of $\sum_{i=1}^{K}(w_{di} \sum_{j=1}^{3}(\alpha_j a_{dij}))$ and variance of δ^2. Thus, we have

$$r_d \sim \mathcal{N}(\sum_{i=1}^{K}(w_{di} \sum_{j=1}^{3}(\alpha_j a_{dij})), \delta^2) \tag{2}$$

As described in Wang et al. [14], different contributors may have different preferences over aspects. For example, contributors may give the same overall rating to an object, but may have different reasons for doing so. Also, one should note that aspects are not independent and an emphasis on one may indicate a preference for the other as well. In order to capture these properties in the model, a Gaussian prior is considered for the aspects weights w_d. Hence,

$$\boldsymbol{w_d} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \tag{3}$$

where $\boldsymbol{\mu}$ is the K-dimensional mean vector and Σ is the $K \times K$ covariance matrix of random vector $\boldsymbol{w_d}$. The probability of overall satisfaction and the preference vector of aspects $\boldsymbol{w_d}$ in each review can be calculated by combining equations

(2) and (3) as

$$\begin{aligned}\Pr(r, \boldsymbol{w}|d) =& \Pr(r_d, \boldsymbol{w_d}|\Theta; \boldsymbol{a_d}) \\ =& p(w_d|\boldsymbol{\mu}, \Sigma)p(r_d|\sum_{i=1}^{K}(w_{di}\sum_{j=1}^{3}(\alpha_j a_{dij})), \delta^2; \boldsymbol{a_d})\end{aligned} \tag{4}$$

$\Theta = (\boldsymbol{\mu}, \Sigma, \delta^2, \boldsymbol{\alpha})$ is the set of model parameters that should be estimated, where $\boldsymbol{\alpha}$ is the 3×1 vector of $[\alpha_1, \alpha_2, \alpha_3]^T$ and $\boldsymbol{a_d}$ is the input aspect information that contains the positive, neutral and negative frequency counts.

4.2. Model parameters estimation

The probabilistic model considered in section 4.1 is modeling the joint distribution of overall rating and linear combination weights $\Pr(r_d, \boldsymbol{w_d})$. In this model the r_d (the overall rating of a review) and $\boldsymbol{a_d}$ (the input aspect information of a review) are observed for each given review. There are four sets of parameters that need to be estimated in the training phase:

- $\boldsymbol{\mu}$ is the mean vector of the prior distribution over the aspects' rating combination coefficients,
- Σ is the covariance matrix,
- δ^2 is the variance of the Normal distribution of the overall satisfaction, and
- $\boldsymbol{\alpha}$ are the mixing coefficients for the positive, neutral and negative sentiment counts of each review.

We denote all parameters together as $\Theta = (\boldsymbol{\mu}, \Sigma, \delta^2, \boldsymbol{\alpha})$.

In equation (4) one can see the preference vector of the aspects $\boldsymbol{w_d}$ as hidden variables and so the model is the joint distribution of some observed and some hidden variables. The Expectation Maximization (EM) algorithm [16] is a well-known algorithm suitable for Maximum Likelihood (ML) parameter estimation with hidden variables. In order to estimate Θ we follow an EM-style algorithm. In the EM algorithm the optimum parameters are those which maximize the log of complete likelihood (observed and hidden variables) of the training data. Thus, the ML estimate $\hat{\Theta}$ of the model parameters is as follows:

$$\hat{\Theta} = arg\max_{\Theta} \sum_{d \in D} \log p(r_d | \boldsymbol{\mu}, \Sigma, \delta^2, \boldsymbol{\alpha}; \boldsymbol{a_d}) \tag{5}$$

In our EM-style algorithm, we first initialize the parameters into Θ_0, and then we alternate between the E-step and M-step in order to update and improve the parameters.

E-step. Using the current parameters estimate, calculate the hidden variables as follows:

$$\text{For each review } d \begin{cases} \text{Calculate the aspect rating using equation (11) .} \\ \text{Calculate } \boldsymbol{w_d} \text{ by solving the problem given in equation (13) below.} \end{cases}$$

M-step. Update the parameters using the calculations done in E-step.

$\boldsymbol{\mu}$ and Σ are the parameters of the Gaussian prior over the linear weights. So, the new estimate for the parameters at the M-step is going to be calculated as the sample mean and sample covariance matrix of $\boldsymbol{w_d}$ computed in the E-step.

$$\boldsymbol{\mu}_{t+1} = \frac{1}{|D|} \sum_{d=1}^{|D|} \boldsymbol{w_d} \tag{6}$$

$$\Sigma_{t+1} = \frac{1}{|D|} \sum_{d=1}^{|D|} (\boldsymbol{w_d} - \boldsymbol{\mu_{t+1}})(\boldsymbol{w_d} - \boldsymbol{\mu_{t+1}})^T \tag{7}$$

δ^2 can be updated in M-step by maximizing the probability defined in equation (2). After solving this optimization problem, the update formula for δ^2_{t+1} will be:

$$\delta^2_{t+1} = \frac{1}{|D|} \sum_{d=1}^{|D|} (r_d - \sum_{i=1}^{K} w_{di} \sum_{j=1}^{3} (\alpha_j a_{dij}))^2 \tag{8}$$

And the updated $\boldsymbol{\alpha}$ would be:

$$\boldsymbol{\alpha_{t+1}} = arg\max_{\boldsymbol{\alpha}} \sum_{d \in D} - \frac{(r_d - \sum_{i=1}^{K} w_{di} \sum_{j=1}^{3} (\alpha_j a_{dij}))^2}{2\delta^2_{t+1}} \tag{9}$$

We define $\boldsymbol{A_d}$ and $\boldsymbol{y_d}$ as

$$\begin{aligned}\boldsymbol{A_d} =& \begin{bmatrix} a_{d_{11}} & a_{d_{12}} & a_{d_{13}} \\ & \dots & \\ a_{d_{K1}} & a_{d_{K2}} & a_{d_{K3}} \end{bmatrix}, \\ \boldsymbol{y_d} =& \boldsymbol{w_d}[1, \quad 1, \quad 1] \odot \boldsymbol{A_d}\end{aligned}$$

where $\odot$ indicates the element-wise product of matrices. Then, the closed-form solution for updating $\boldsymbol{\alpha}$ can be worked out as:

$$\boldsymbol{\alpha_{t+1}} = [\sum_{d \in D} \boldsymbol{y_d} \boldsymbol{w_d}^T \boldsymbol{A_d}]^{-1} \sum_{dinD} r_d \boldsymbol{y_d} \tag{10}$$

The E and M steps are repeated until convergence of the algorithm.

4.3. How to use the learned probabilistic model

In this section we will describe, given the model parameters $\Theta = (\boldsymbol{\mu}, \Sigma, \delta^2, \boldsymbol{\alpha})$, how one can use the proposed model in order to calculate the aspect rating for each individual aspect and derive the relative importance of the aspects from the contributor's perspective. The aspect rating for the i^{th} aspect in review d can be calculated as

$$a_{di} = \sum_{j=1}^{3} \alpha_j a_{dij} \tag{11}$$

using the learned $\boldsymbol{\alpha}$.

In the proposed model in section 4.1 the linear combination weights $\boldsymbol{w_d}$ show the relative importance of each aspect in each review from the contributors' point of view. For a given review (given r_d and $\boldsymbol{a_d}$) and model parameters (given Θ), it is possible to find the most probable linear combination of the weights $\boldsymbol{w_d}$. One way to do that is to find those $\boldsymbol{w_d}$ which maximize the posterior probability of the given review (Maximum A Posteriori (MAP) estimation). The objective function of the MAP estimation is the log of the joint distribution of $\Pr(r_d, \boldsymbol{w_d})$, where

$$f_{MAP}(d) = \log p(w_d | \boldsymbol{\mu}, \Sigma) p(r_d | \sum_{i=1}^{K} (w_{di} \sum_{j=1}^{3} (\alpha_j a_{dij})), \delta^2) \tag{12}$$

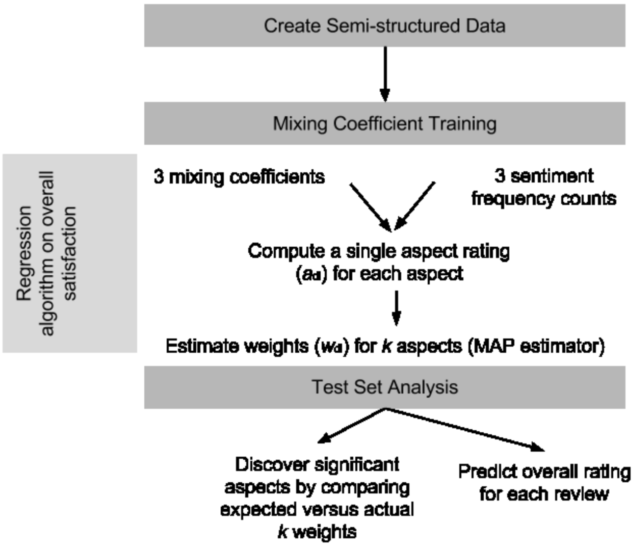

Create Semi-structured Data
Mixing Coefficient Training
Regression algorithm on overall satisfaction
3 mixing coefficients
3 sentiment frequency counts
Compute a single aspect rating (a_d) for each aspect
Estimate weights (w_d) for k aspects (MAP estimator)
Test Set Analysis
Discover significant aspects by comparing expected versus actual k weights
Predict overall rating for each review

Figure 3: Overview of the methodology.

and therefore the MAP estimator of importance weights $\hat{\boldsymbol{w}}_{\boldsymbol{d}}$ is

$$\begin{aligned}\hat{\boldsymbol{w}}_{\boldsymbol{d}} =& \arg\max f_{MAP}(d) \\ =& \arg\max[-\frac{(r_d - \boldsymbol{w_d}^T\boldsymbol{a_d})^2}{2\delta^2} - \frac{1}{2}(\boldsymbol{w_d} - \boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{w_d} - \boldsymbol{\mu})]\end{aligned} \tag{13}$$

subject to the following constraints:

$$\sum_{i=1}^{K} w_{di} = 1$$
$$0 \leq w_{di} \leq 1 \quad \text{for} \quad i = 1, 2, \ldots, K.$$

The optimization problem in equation (13) is a quadratic programming problem and can be solved using the standard (existing) methods for such optimization problems. A sequential least squared programming method was used to solve this problem. A high level overview of our combined methodology is depicted in Figure 3. Starting with the creation of the semi-structured data, as depicted in Figure 1, we now see the overall process showing the computation of aspect ratings and MAP estimation of weights. Section 5 will discuss the analysis performed using test data from *TripAdvisor.com*.

5. Experimental Results

In this section the experimental data set is described, and we then explain the details of the experiments conducted.

Data set. We used reviews that visitors have put on *TripAdvisor.com* to create our corpus. The reviews were collected from 51 different state parks in the state of California. In order to select the parameters of the model, we have divided our data into training and testing sets. The training data is used for parameter estimation and the test set is used for evaluation of the methodology.

We used quota sampling by randomly selecting 17 parks from three predetermined park categories: beach, historic, and nature. These park categories are only used for quota sampling in order to capture the service diversity within the data set. For each park category, we randomly selected 10 of the 17 parks for training and 7 for the test set, for a total of 30 parks in the training set, and 21 in the test set. The entire training set consists of $7,888$ positive, $4,120$ neutral and $1,458$ negative sentences, for a total of $13,466$ sentences overall. The entire test set has $1,803$ positive, $1,026$ neutral and 460 negative sentences ($3,289$ sentences overall). After transforming all text to lower cases, the word tokens are augmented with part-of-speech (POS) tags and then stop words are removed. Terms which occur fewer than 10 times are discarded and the stems of the remaining tokens are used to construct the word list. Table 1 gives more details about the data set.

Aspect identification. Using the BON representation of the sentences in the training set and the aspect identification subsystem (Figure A.4 (a)) the aspects that the users have expressed their opinions about were identified as **"beach", "camp", "hike", "history", "nature", "park", "ranger", "road", "shop", "shower", "trail", "tour" and "view".** These $K = 13$

Table 1: Data set statistics

	Training	Test	Overall
Number of parks	30	21	51
Number of reviews	1,791	448	2,239
Number of sentences	13,466	3,289	16,755

Table 2: ML Estimated of mixing coefficients.

α_1 (for positive sentiments)	α_2 (for neutral sentiments)	α_3 (for negative sentiments)
20.1326	12.9061	14.6074

aspects have been identified as the representative terms of clusters. These aspects were then used in the rest of the experiments in order to give structure to the unstructured reviews.

Give structure to data and learn the model. For each given review and for each of the identified aspects, the number of sentences with positive, neutral and negative sentiments are counted and normalized separately. The ground truth sentiment of the sentences in the training set are given manually. We have used the Delphi method with three human evaluators to determine the ground truth consensus. In the test phase, it is possible to use the sentiment identification subsystem (Figure A.4 (b)) to predict the sentiment of each sentence which contains one of the identified aspects. Using the described scheme, each review in the training and test set is transformed into a 39-dimensional (3×13) vector. The parameters ($\boldsymbol{\Theta}$) of the model proposed in section 4.1 were estimated using this structured data and the EM-style algorithm described in section 4.2. The mixing coefficients are used to get a single aspect rating for each aspect. The estimated mixing coefficients ($\boldsymbol{\alpha}$) are shown in Table 2. It can be seen that in calculating a single rating for each aspect, our Bayesian framework puts more weight on positive and negative sentiments than neutral ones. Also, the weight of positive sentiments is more than the negative ones.

Table 3: Aspect importance using ground truth sentiments.

Test park #	**Beach**	**Camp**	**Hike**	**History**	**Nature**	**Park**	**Ranger**	**Road**	**Shop**	**Shower**	**Trail**	**Tour**	**View**
1	0.06854	**0.16711**	**0.07501**	0.06601	0.06523	0.05703	**0.10531**	0.05736	0.07483	0.04984	0.05287	0.05851	**0.10235**
2	**0.16780**	0	**0.15154**	0	0	**0.11721**	0.06755	0	**0.38667**	0.01146	0	0	**0.09776**
3	0.04964	0.07117	**0.09431**	**0.09908**	0.07391	0.06426	**0.08473**	0.06299	0.0762	0.05377	**0.10709**	0.06816	**0.09469**
4	0.02975	0.03076	0.06617	**0.24449**	**0.1345**	0.05226	**0.0827**	0.02876	0.051	0.04814	**0.12603**	0.02502	**0.08044**
5	0.01272	0.00784	**0.17289**	**0.09766**	0.00754	**0.13984**	0.06573	0.00926	0.04382	**0.07618**	0.00836	**0.27037**	**0.08782**
6	0.02019	0.0369	0.01776	0.00203	**0.12903**	0.02451	**0.08**	0.0144	0.01200	0.04833	**0.297**	**0.23142**	**0.0865**
7	0.01513	0.01323	0.07175	**0.24644**	**0.16251**	0.01423	0.059	**0.1735**	0.02701	0.06170	0.00866	0.0495	**0.09739**
8	0.0721	**0.25708**	0.04696	0.05469	0.06721	0.04701	**0.12725**	0.04112	0.03512	0.04987	0.05659	0.03950	**0.10548**
9	0.0352	0.03825	**0.08665**	**0.25090**	**0.11918**	0.06350	**0.08019**	0.04376	0.06830	0.05053	0.03938	0.03962	**0.08457**
10	0.05266	**0.07661**	**0.09505**	**0.07665**	0.06857	0.07030	**0.08351**	0.04956	**0.08856**	0.05202	0.05456	**0.13584**	**0.09610**
11	0.0518	**0.10176**	**0.08040**	**0.23005**	**0.10207**	0.07029	**0.09735**	0.01514	**0.07755**	0.04612	0.017	0.02374	**0.08675**
12	**0.12267**	0	0	0	0.1644	0.00648	0.0478	0	**0.28654**	0	0	**0.273**	**0.09911**
13	**0.09258**	**0.37424**	0.01126	0.02362	0.06832	0.03172	**0.15434**	0.02607	0.02488	0.04327	0.01838	0.01873	**0.11259**
14	0.03924	0.01871	0.06006	**0.32673**	**0.16236**	0.05012	**0.08021**	0.01764	**0.09201**	0.03801	0.01909	0.01838	**0.07745**
15	0.05389	0.02816	0.04387	0.03567	**0.12593**	0.03655	0.06888	0.03458	**0.12432**	0.03562	0.02943	**0.28867**	**0.09445**
16	**0.07868**	**0.17650**	0.07031	0.05894	0.06548	0.05531	**0.10841**	0.05074	**0.09509**	0.04485	0.04238	0.05032	**0.10299**
17	**0.08006**	**0.19414**	0.04314	0.03607	**0.08313**	0.04334	**0.11181**	0.03162	**0.09282**	0.03917	0.03027	**0.11125**	**0.10314**
18	0.06538	0.05459	**0.10385**	0.04463	0.06359	0.05515	**0.07721**	**0.10724**	**0.11741**	0.05122	**0.11195**	0.04605	**0.10174**
19	0.06667	**0.18113**	**0.07811**	**0.09793**	0.06325	0.06508	**0.11012**	0.04160	0.06648	0.05074	0.03077	0.04785	**0.10027**
20	0.05997	**0.22014**	0.03022	**0.10810**	**0.10394**	0.03826	**0.11860**	0.02781	0.03174	0.0460	0.02747	**0.08797**	**0.09977**
21	0.06356	**0.14984**	0.04899	**0.07503**	**0.09796**	0.03996	**0.10063**	0.05112	0.07422	0.04428	0.04894	**0.10537**	**0.10012**

Important aspects from contributors' perspectives. The purpose of this set of experiments is to determine the relative importance of the identified aspects for each park from the park's visitors' perspectives. As described in section 4.1 in the proposed probabilistic model, the $\boldsymbol{w_d}$ indicates the relative importance of each aspect in the review. Given the model parameters $\boldsymbol{\Theta}$, the relative importance vector can be estimated for each review using the MAP estimator described in section 4.3. Here it should be noted that for each park in the test set, the sentiment information vector of each review $\boldsymbol{a_d}$ can be generated using either the ground truth sentiment labels or the predicted sentiment labels obtained by the sentiment identification subsystem (Figure A.4 (b)). Tables 3 and 4 show the results of these experiments, using the ground truth sentiment labels and predicted sentiment labels respectively.

In these experiments, the sentiment information vector $\boldsymbol{a_d}$ was computed for each review using the mixing coefficients $\boldsymbol{\alpha}$ learned from the training data. These sentiment information vectors were then averaged over all reviews of each test park ($\boldsymbol{\bar{a}_d}$). Next, the relative importance vector was estimated by the MAP estimator described in section 4.3 for the given test park using $\boldsymbol{\bar{a}_d}$. In

Table 4: Aspect importance using predicted sentiments.

Test park #	Beach	Camp	Hike	History	Nature	Park	Ranger	Road	Shop	Shower	Trail	Tour	View
1	0.06647	**0.11719**	**0.09745**	0.05638	0.05636	0.06040	**0.09176**	**0.08894**	**0.09225**	0.05329	0.06995	0.04619	**0.10336**
2	**0.16911**	0	**0.14934**	0	0	**0.12321**	0.06396	0	**0.38617**	0.01143	0	0	**0.09676**
3	0.06661	0.05977	**0.12324**	**0.08691**	0.04596	**0.08311**	**0.08189**	0.06326	**0.12716**	0.05111	0.05954	0.05476	**0.09669**
4	0.03660	0.03130	0.07459	**0.17366**	**0.11148**	0.05801	**0.08147**	0.02938	0.06549	0.04857	**0.14506**	0.06050	**0.08389**
5	0.02551	0.01053	**0.16532**	**0.09011**	0.01444	**0.12865**	0.06667	0.02341	0.06987	0.07033	0.01160	**0.23348**	**0.09008**
6	0.02455	0.03885	0.02429	0.00671	**0.11728**	0.02991	**0.08564**	0.00630	0.01322	0.04822	**0.36343**	**0.15758**	**0.08401**
7	0.03781	0.02803	0.05980	**0.13169**	**0.14107**	0.02055	0.06435	**0.12442**	**0.07648**	0.04986	0.02543	**0.14164**	**0.09887**
8	0.07117	**0.24194**	0.05095	0.03417	0.06393	0.05019	**0.12192**s	0.03961	0.04392	0.04983	0.04021	**0.08619**	**0.10598**
9	0.03137	0.02964	**0.08162**	**0.29391**	**0.13363**	0.06246	**0.07989**	0.03387	0.06565	0.04880	0.02889	0.02925	**0.08103**
10	0.05198	**0.08459**	**0.09134**	0.07080	0.06721	0.07480	**0.08592**	0.02936	**0.08654**	0.05094	0.04143	**0.17019**	**0.09489**
11	0.03269	0.05527	**0.09054**	**0.26035**	**0.11104**	0.07308	**0.08500**	0.02649	0.05663	0.05258	0.02394	0.04895	**0.08344**
12	0	0	0	0	**0.17179**	0.02453	0.03663	0	0.00915	0.05199	0	**0.61243**	**0.09347**
13	**0.09336**	**0.32923**	0.02558	0.03212	0.06704	0.03703	**0.14412**	0.03274	0.05035	0.04211	0.01793	0.01763	**0.11077**
14	0	0	0	**0.52977**	**0.26704**	0.01473	**0.07620**	0	0.01310	0.03566	0	0	0.06351
15	0.05398	0.03651	0.04689	0.04147	**0.12133**	0.03926	0.07152	0.03375	**0.11929**	0.03689	0.03408	**0.27064**	**0.09440**
16	0.06300	**0.15706**	0.07110	0.07061	0.07117	0.05536	**0.10379**	0.05267	0.06564	0.05037	0.07295	0.06587	**0.10040**
17	0.07485	**0.15336**	0.06265	0.06409	**0.07930**	0.05123	**0.10311**	0.04320	**0.09980**	0.04220	0.04253	**0.08307**	**0.10061**
18	0.04069	0.03634	**0.13446**	0.03462	0.05277	0.04635	0.06043	**0.21390**	0.06193	0.07357	**0.09569**	0.03669	**0.11255**
19	**0.08601**	0.03928	**0.10999**	0.05068	0.05486	0.07340	**0.07912**	0.05336	**0.18299**	0.03815	0.07363	0.06194	**0.09658**
20	0.06617	**0.20466**	0.04404	0.05372	**0.08443**	0.04304	**0.11314**	0.04067	0.05269	0.04676	0.04062	**0.10702**	**0.10303**
21	0.06513	**0.11917**	0.06942	**0.09824**	**0.08981**	0.04929	**0.09438**	0.06056	**0.09318**	0.04526	0.04666	0.07019	**0.09870**

Tables 3 and 4 those aspects for which their relative importance is greater than the uniform aspect weight have been considered as important aspects for each test park (bold numbers). According to Table 3, for instance, the significant aspects of Park 1 are: “Camp”, “Hike”, “Ranger” and “View” from its visitors’ perspective. The managers can get valuable insights from their customers about the different aspects of their service, and realize the strengths and weaknesses of the park’s services. Also, potential visitors can now take their peers’ opinions into account when deciding whether to visit a park.

To quantify the error rate for identifying the significant aspects appropriate to each park, we performed a diagnostic test using binary classification. In this experiment we treated bold numbers in Tables 3 and 4 as “positive” and unbolded text as “negative”. Also we considered the results given using the ground truth data (Table 3) as ground truth and evaluated the performance of the ML estimation of significant aspects using the predicted labels (Table 4). Table 5 summarizes the results of this diagnostic test, and the details of the computations for the diagnostic variables are given here:

$$\begin{aligned}
\text{Accuracy} &= \frac{\sum \text{True Positive} + \sum \text{True Negative}}{\sum \text{Total Population}} \\
\text{Precision} &= \frac{\sum \text{True Positive}}{\sum \text{Test Outcome Positive}} \\
\text{Negative Predictive Value} &= \frac{\sum \text{True negative}}{\sum \text{Test Outcome Negative}} \\
\text{Sensitivity} &= \frac{\sum \text{True Positive}}{\sum \text{Ground Truth Positive}} \\
\text{Specificity} &= \frac{\sum \text{True Negative}}{\sum \text{Ground Truth Negative}}
\end{aligned}$$

where

$$\sum \text{Test Outcome Positive} = \sum \text{True Positive} + \sum \text{False Positive}$$

$$\sum \text{Test Outcome Negative} = \sum \text{True Negative} + \sum \text{False Negative}$$

$$\sum \text{Ground Truth Positive} = \sum \text{True Positive} + \sum \text{False Negative}$$

$$\sum \text{Ground Truth Negative} = \sum \text{False Positive} + \sum \text{True Negative}$$

As shown in Table 5, if we make use of sentiment classification to obtain positive, neutral, and negative sentiments associated with each aspect in the test set, as opposed to using the ground truth sentiments, we only notice a slight deterioration in our ability to appropriately categorize the significant aspects. If no deterioration is allowed, then the time and expense associated with obtaining ground truth sentence fragment sentiments would be required, but would likely be impractical for most real data sets. For example, *TripAdvisor* contains more than 150 million reviews as of the first quarter of 2014 (TripAdvisor.com). An average overall accuracy of 88.3% clearly demonstrates the ability determine the significant aspects associated with overall customer satisfaction.

Next, using the model obtained from the training set, we apply the weightings to the positive, neutral, and negative comments for all aspects to obtain a

Table 5: Diagnostic test results.

	Average	Minimum	Std. dev.
Accuracy	0.883	0.652	0.090
Precision	0.890	0.500	0.170
Negative predictive value	0.888	0.750	0.091
Sensitivity	0.819	0.500	0.162
Specificity	0.934	0.667	0.106

predicted value rounded to the nearest integer for overall customer satisfaction for each contributor for a particular park. The accuracy of our model to predict the customers self-reported overall customer satisfaction on a 5-point scale is presented in Table 7 as R^2. Additionally, we simplify the overall customer satisfaction to a binary scale, with overall customer satisfaction scores of 1, 2, and 3 being labeled 0, and scores of 4 and 5 being labeled 1. This common technique converts the 5-point Likert scale into a 2-point binary scale representing bad versus good overall customer satisfaction. The accuracy of our model to predict binary overall customer satisfaction on a 5-point scale is presented in Table 7 as R^2 (binary).

We obtain the R^2 values for both the ground truth and predicted sentiment labels. The R^2 and R^2 (binary) values are presented for both the ground truth and predicted labels, and the formula is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y_i})^2}{\sum_{i=1}^{N}(y_i - \bar{y_i})^2}$$

where N is the number of usable reviews, y_i is self-reported overall customer satisfaction for a contributor, $\hat{y_i}$ is the overall customer satisfaction predicted by the model for that customer, and $\bar{y}$ is the mean of the y_i values. We omit reviews containing either none or only one sentence fragment mentioning any identified aspect. These are extremely sparse reviews in terms of information relating to the aspects deemed relevant in the training set. We justify exclusion based on the same principal used to exclude incomplete responses in traditional surveys because responses with zero or one sentence fragment mentioning any aspect provide no evaluative trade-off. Presented in Table 6 are the total number of reviews for each park, as well as the number of these reviews that were used in our R^2 analysis. We have omitted reviews that are too short and cover fewer than 2 aspects in our analysis. In Table 6 the distribution of the given overall satisfaction for all/usable reviews have been shown. As seen in Table 7, the average R^2 value for the ground truth is 0.910, and the average R^2 value for the predicted labels is 0.892. The average R^2 (binary) is 0.999 for both the ground truth and the predicted labels.

6. Extended Experiments

In order to test our proposed framework more extensively, we performed an additional set of experiments. In these experiments we used the data from an additional 102 state parks. The ground truth labels were not available for these 102 data sets. We predicted the labels of each sentence using the sentiment identification subsystem (Figure A.4), and using these labels, we generated the survey-like data (a $3K$–dimensional vector) for each review. Having this vector and our Bayesian framework, the overall rating of each park was estimated and compared with the actual overall rating. The R^2 analysis of these experiments are reported in Table 8.

For six of these 102 parks, a negative R^2 is produced, which is a clear indication that the prediction model is not a good fit for the given data. Either the comments are insufficient to extract meaningful information, or the aspects relevant to these parks are outside the relevant range of the training set. Examination in detail of individual comments for these 6 parks reveals that most of the reviews were very short in length. Additionally, the contributors usually mentioned fewer than 2 aspects discovered in our training set. Therefore, we conclude that these 6 parks are either outside the relevant range of the training set or that the reviews are insufficient to rely upon the results. In the remaining

Table 6: Number of All/Usable Reviews by Park and Rating.

		All/Usable Reviews by Rating				
Park	All/Usable reviews	1	2	3	4	5
1	44/34	0/0	1/1	1/1	16/9	26/23
2	10/4	0/0	0/0	0/0	3/2	7/2
3	39/28	1/0	1/1	5/4	10/6	23/17
4	22/20	1/1	0/0	3/3	8/3	10/8
5	10/10	2/2	2/2	0/0	2/2	4/4
6	22/16	4/1	7/6	0/0	10/8	1/1
7	17/15	0/0	1/0	0/0	4/4	12/11
8	28/21	0/0	0/0	0/0	7/6	21/15
9	20/4	0/0	1/0	3/1	5/2	11/1
10	37/21	0/0	0/0	2/0	10/7	25/14
11	18/6	0/0	1/0	5/1	6/3	6/2
12	10/2	0/0	0/0	0/0	5/1	5/1
13	27/20	0/0	0/0	1/1	16/14	10/5
14	22/6	1/0	1/0	1/0	5/2	14/4
15	13/13	0/0	1/1	1/1	3/3	8/0
16	31/24	0/0	2/1	1/1	7/6	21/16
17	12/12	0/0	0/0	0/0	7/7	5/5
18	14/8	0/0	0/0	0/0	3/2	11/6
19	18/12	0/0	0/0	0/0	7/4	11/8
20	27/22	0/0	1/1	1/0	10/8	15/13
21	21/17	0/0	0/0	2/0	7/6	12/11

Table 7: R^2 analysis for test parks.

	Ground truth		**Predicted labels**	
Park	R^2	R^2 (binary)	R^2	R^2 (binary)
1	1.0	1.0	1.0	1.0
2	1.0	1.0	1.0	1.0
3	0.855	1.00	0.613	0.995
4	0.850	0.996	0.80	0.996
5	1.0	1.0	0.924	0.988
6	0.912	1.0	0.912	1.0
7	0.692	1.0	0.692	1.0
8	1.0	1.0	1.0	1.0
9	1.0	1.0	1.0	1.0
10	1.0	1.0	1.0	1.0
11	1.0	1.0	1.0	1.0
12	1.0	1.0	1.0	1.0
13	1.0	0.996	1.0	0.996
14	1.0	1.0	1.0	1.0
15	1.0	1.0	1.0	1.0
16	0.785	0.997	0.785	0.997
17	1.0	1.0	1.0	1.0
18	1.0	1.0	1.0	1.0
19	1.0	1.0	1.0	1.0
20	0.914	1.0	0.913	1.0
21	1.0	1.0	1.0	1.0
average	**0.91**	**0.999**	**0.892**	**0.999**

96 of the 102 additional parks tested, the average R^2 is 0.829. Thus, we conclude that our predictive model is acceptable for use for 96 of the 102 additional parks tested, explaining on average 82.9% of the variation in the overall rating, and 99.8% for the binary model.

7. Discussion

This work describes a method to discern the significant aspects of customer satisfaction latent in free-form customer reviews, such as those found in the TripAdvisor travel destination website. Utilizing semi-structured data, this paper proposes a Bayesian approach to model the overall customer satisfaction in terms of the aspects identified from the nouns in the text and the positive, neutral, and negative sentiments associated with each aspect. This Bayesian model considers the overall rating of each review as a weighted sum of the ratings of the individual aspects. This model enables us to estimate a single rating for each discovered aspect from each contributors perspective. Also, using the model, it is possible to infer (using MAP estimation) the relative significance of each aspect from the contributors point of view.

There are many settings for which it is not possible to obtain the overall rating (e.g., omitted data and verbal complaints where it is not appropriate to ask for an overall rating). In such situations, our approach can be utilized to predict the overall rating for these individuals.

Empirical experiments were conducted using TripAdvisor data for California State Parks. Ground truth sentiment data regarding positive, neutral, or negative sentiment of sentence fragments was obtained using a three-person Delphi method, where at least two out of three evaluators are required to be in agreement to establish the ground truth. The cluster analysis, sentiment classification and probabilistic Bayesian modeling of overall customer satisfaction is applied to the ground truth data. The significant aspects of overall customer satisfaction for a test set of 21 California State Parks is found using this ground truth sentiment evaluation. A comparison is made to the significant aspects of

Table 8: R^2 analysis for extended experiments

Park	R^2	R^2 (binary)	**Park**	R^2	R^2 (binary)	**Park**	R^2	R^2 (binary)	**Park**	R^2	R^2 (binary)
1	0.747	0.994	27	0.799	1.0	53	1.0	1.0	79	0.522	0.993
2	0.767	0.999	28	0.830	1.0	54	1.0	1.0	80	1.0	1.0
3	0.788	0.998	29	0.759	0.997	55	0.889	0.999	81	1.0	1.0
4	0.821	0.999	30	0.883	0.994	56	1.0	1.0	82	1.0	1.0
5	0.480	1.0	31	1.0	1.0	57	0.929	1.0	83	1.0	1.0
6	0.839	1.0	32	0.926	1.0	58	0.639	0.999	84	0.956	1.0
7	0.919	1.0	33	-7.33	0.967	59	0.845	0.992	85	0.784	0.996
8	0.854	1.0	34	0.555	0.993	60	0.458	1.0	86	1.0	1.0
9	0.359	1.0	35	0.879	0.981	61	0.639	0.992	87	1.0	1.0
10	1.0	1.0	36	0.921	1.0	62	0.579	1.0	88	0.746	0.997
11	1.0	1.0	37	0.816	1.0	63	-0.377	0.995	89	0.949	0.998
12	0.858	1.0	38	0.691	0.995	64	1.0	1.0	90	1.0	1.0
13	0.892	1.0	39	1.0	1.0	65	0.30	0.995	91	0.778	1.0
14	0.862	1.0	40	0.767	1.0	66	0.332	0.989	92	1.0	1.0
15	0.934	1.0	41	0.881	0.998	67	1.0	1.0	93	0.851	0.999
16	0.813	0.998	42	1.0	1.0	68	0.649	1.0	94	1.0	1.0
17	0.702	1.0	43	1.0	1.0	69	1.0	1.0	95	0.919	1.0
18	-4.257	0.988	44	1.0	1.0	70	1.0	1.0	96	0.885	1.0
19	0.890	1.0	45	0.531	0.996	71	1.0	1.0	97	0.813	1.0
20	0.818	0.995	46	-0.302	0.988	72	1.0	1.0	98	1.0	1.0
21	-1.0	1.0	47	0.125	0.994	73	0.803	1.0	99	1.0	1.0
22	0.643	0.989	48	0.531	0.987	74	1.0	1.0	100	1.0	1.0
23	0.409	1.0	49	0.842	0.998	75	1.0	1.0	101	0.445	0.989
24	0.840	1.0	50	0.966	1.0	76	1.0	1.0	102	1.0	1.0
25	0.736	0.993	51	0.334	0.995	77	1.0	1.0			
26	0.918	1.0	52	-1.246	1.0	78	0.928	1.0			

customer satisfaction obtained using the sentiment classification method, with an average overall accuracy of 88.3%. This result means that sentiment classification with Bayesian parameter estimation is almost as accurate as using the ground truth to determine the significant aspects that drive customer satisfaction. Average R^2 values for predicting the overall customer satisfaction rating ranged from 0.892 using predicted labels and 5-point overall customer satisfaction ratings to 0.999 using ground truth and a binary overall customer satisfaction classification.

Obtaining results from the training set requires extensive manual labor in order to establish the ground truth. Once the training results are obtained, however, the application of the results from the training set to the testing set is very scalable. The average computational time is 0.03285 seconds per product or service (in this case, per park) using a MacBook Pro with a 3.2 GHz Intel i5 four core CPU and 8 Gb of RAM. The minimum computational time was 0.02535 seconds, the maximum was 0.04715 seconds, and the standard deviation was 0.00416 seconds. Hypothetically, in 24 hours, 2.63 million distinct products or services could be analyzed using our method.

To apply this methodology effectively, some additional insight may be helpful. First, the training set aspects must be sufficiently similar to the true set of aspects latent in the test set. If the true aspects latent in the reviews for a particular object in the test set are different from the aspects discovered in the training set, then the test set may be outside the relevant range of analysis for the given training set. Indications that additional aspects may also be relevant for a particular test set object are either a very low R^2 value or an unusually large number of omitted reviews. A related insight is that the training set must be sufficiently robust to be relevant to the test set. In our case, using 30 parks for the training set, about 1800 reviews, was sufficient. Additionally, there is no previously established test to determine if the test object is outside the relevant range of the training set. Two indicators that the test object is within then relevant range of the training set are high R^2 values for prediction of individual overall customer satisfaction scores and a reasonably low number of omitted reviews. Standard conventions related to strong R^2 values appear to be reasonable, with values above 0.8 considered extremely strong, values between 0.4 and 0.8 being reasonably strong, values between 0.2 and 0.4 being acceptable, and below 0.2 considered weak. Omitted reviews can be examined to determine whether omission occurred because other aspects are present in the reviews and the object is truly outside the relevant range, or because the reviews did not disclose underlying aspects driving the contributors' overall customer satisfaction.

The second insight is to be cognizant of potential biases in the data. We acknowledge that some of the TripAdvisor reviews may be falsified by management to make their service look better to either customers or to their superiors. In the case of California State Parks, we speculate that the pressure on management to falsely write reviews will be substantially less than it is for commercial enterprises. Additionally, TripAdvisor does purport to make attempts to remove falsified reviews. This is an issue with all online reviews and has been addressed in previous work by Mukherjee, Liu and Glance [17].

8. Conclusions

The major contribution of this paper is the linking of aspect identification and semantic classification methods to explain and predict overall customer satisfaction. First, a method is proposed by which unstructured user generated text data is transformed into ready-to-analyze data without the need to determine aspects a priori. Second, a Bayesian model is proposed that allows prediction of individual aspect ratings, and further enables discovery of the relative importance of each aspect from each contributor's perspective. Consequently, the method also allows for prediction of overall customer satisfaction. The model presented in this paper has low dimensionality that can be scaled to analyze very large data sets in an automated fashion. Results of the Bayesian method are reproducible. Thorough testing illustrates that the methods presented in this paper are effective in discovering, explaining, and predicting the most important aspects driving overall customer satisfaction.

The proposed method has several very important applications. First, managers can use our methodology to assess the importance of aspects that drive the overall customer satisfaction. Secondly, the methodology is capable of enabling recommendation systems to match aspect evaluations to a prospective customer's preference profile. Third, traditional survey systems may be augmented with these methods to enable the discovery of the aspects of a product or service that influence overall customer satisfaction, rather than pre-specifying the aspects, as currently occurs with traditional survey methods. The amount of data that market researchers are able to evaluate using this method would be substantially larger than with traditional surveys, and the cost to do so would be significantly lower, all due to automation. Fourth, the model can be used to predict a particular customer's overall satisfaction, given that they provide a free-form text review. In situations such as a telephone or online complaint system where the overall satisfaction is not practical to obtain, the customer's overall satisfaction scores can be estimated. Finally, we posit that our method can be used to monitor changes in the major drivers, or aspects, of customer satisfaction over time by comparing aspects across different time-periods. This means that the methodology could, in effect, enable us to monitor changes in satisfaction drivers over time and alert the product or service provider to changing customer preferences.

There are many opportunities for extending this research. Review engines often have unique editorial restrictions, which may have impacts that are not well understood. For example, TripAdvisor enforces a minimum of 200 characters in a review. One interesting question would be to develop guidance or policies as to setting such restrictions, and how these policies would affect the ability to extract aspects and determine their significance on overall satisfaction. Aspect information, as well as their significance, may possibly also be aided by taking into account reviewer information such as the reviewer's editorial role, how many reviews they have posted reviews, and other readily available metadata. Finally, it would be interesting to understand under which conditions early reviews may impact the aspects and aspect ratings of later reviews for the same product or service. More research is needed in all of these areas, as they are important to improving our understanding of customer satisfaction, as well as improving product and service review systems.

Acknowledgment

We wish to thank Dr. Roummel Marcia, University of California, Merced and Dr. Paul Messinger, University of Alberta for their helpful suggestions.

References

[1] J. Merkert, M. Mueller, M. Hubl, A survey of the application of machine learning in decision support systems, in: Twenty-Third European Conference on Information Systems (ECIS), ACIS, 2015.

[2] K. F. Ward, E. Rolland, R. A. Patterson, Improving outpatient health care quality: understanding the quality dimensions, Health Care Management Review 30 (4) (2005) 361–371.

[3] M. Farhadloo, E. Rolland, Multi-class sentiment analysis with clustering and score representation, in: IEEE 13th International Conference on Data Mining, Workshops (ICDMW), IEEE, 2013, pp. 904–912.

[4] S. Moghaddam, M. Ester, Opinion digger: an unsupervised opinion miner from unstructured product reviews, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, ACM, 2010, pp. 1825–1828.

[5] D. M. Blei, Probabilistic topic models, Communications of the ACM 55 (4) (2012) 77–84.

[6] Y. Jo, A. H. Oh, Aspect and sentiment unification model for online review analysis, in: Proceedings of the fourth ACM international conference on Web search and data mining, WSDM '11, ACM, 2011, pp. 815–824.

[7] H. Lakkaraju, C. Bhattacharyya, I. Bhattacharya, S. Merugu, Exploiting coherence for the simultaneous discovery of latent facets and associated sentiments, in: Proceedings of SIAM International Conference on Data Mining (SDM), SIAM, 2011, pp. 498–509.

[8] S. Moghaddam, M. Ester, On the design of lda models for aspect-based opinion mining, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, ACM, 2012, pp. 803–812.

[9] F. Xianghua, L. Guo, G. Yanyan, Z. Wang, Multi-aspect sentiment analysis for chinese online social reviews based on topic modeling and hownet lexicon, Knowledge-Based Systems 37 (2013) 186–195.

[10] Z. Gan, C. Chen, R. Henao, D. Carlson, L. Carin, Scalable deep poisson factor analysis for topic modeling, in: Proceedings of the 32nd International Conference on Machine Learning (ICML-15), 2015, pp. 1823–1832.

[11] M. Zhou, L. Carin, Negative binomial process count and mixture modeling, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (2) (2015) 307–320.

[12] E. Cambria, S. Poria, F. Bisio, R. Bajpai, I. Chaturvedi, The clsa model: a novel framework for concept-level sentiment analysis, in: Computational Linguistics and Intelligent Text Processing, Springer, 2015, pp. 3–22.

[13] M. Farhadloo, E. Rolland, Fundamentals of sentiment analysis and its applications, in: W. Pedrycz, S. Chen (Eds.), Sentiment Analysis and Ontology Engineering: An Environment of Computational Intelligence, Vol. 639 of Studies in Computational Intelligence, Springer, 2016, pp. 1–24.

[14] H. Wang, Y. Lu, C. Zhai, Latent aspect rating analysis on review text data: a rating regression approach, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2010, pp. 783–792.

[15] B. Liu, Sentiment analysis and subjectivity, in: Handbook of natural language processing, Vol. 2, Chapman & Hall, 2010, pp. 627–666.

[16] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the *em* algorithm, Journal of the Royal Statistical Society 39 (1) (1977) 1–38.

[17] A. Mukherjee, B. Liu, N. Glance, Spotting fake reviewer groups in consumer reviews, in: Proceedings of the 21st International Conference on World Wide Web, WWW '12, ACM, 2012, pp. 191–200.

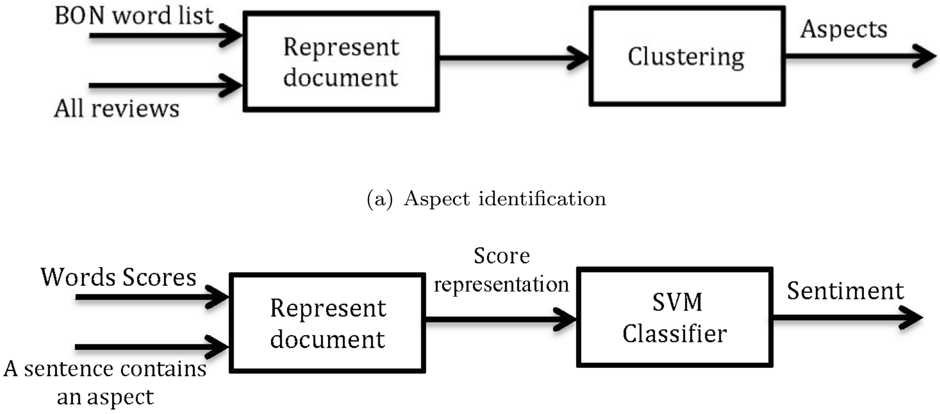
[18] M. Gamon, A. Aue, S. Corston-Oliver, E. Ringger, Pulse: Mining customer opinions from free text, in: Advances in Intelligent Data Analysis VI, Springer, 2005, pp. 121–132.

Appendix A. Transforming unstructured data to semi-structure data

Farhadloo and Rolland [3] proposed a new framework for aspect-level sentiment analysis. In aspect level sentiment analysis, the aspects of the object that the user has commented on is first identified, and then the sentiment of the sentence about that aspect is discovered. For the aspect identification step, the authors proposed to not ignore the part-of-speech tags, and instead of clustering with bag of words, employ a clustering over the sentences only using bag of nouns [3]. They also proposed a new feature set, score representation, that led to more accurate sentiment identification. That scheme is based upon the three scores (positiveness, neutralness and negativeness) that are learned from data for each term in the vocabulary list. The framework in [3] consists of two stages as shown in Figure A.4.

The first step in aspect level sentiment analysis is to identify the aspects that the users have expressed their opinion about in the sentences. For this purpose, Farhadloo and Rolland [3] followed the idea of employing clustering over sentences in order to identify the aspects. However instead of using Bag of Words (BOW) for clustering which doesn't produce satisfactory results [18], they proposed to use Bag Of Nouns (BON). Their results showed that BON representation improved the performance of clustering and made clustering an effective algorithm for aspect identification (Figure A.4 (a)). For sentiment identification, they followed a machine learning approach by designing a 3-class SVM classifier. However, instead of using the commonly used BOW representation, usually with high dimension, they proposed a new feature set based on positiveness, neutralness and negativeness scores (a 3-dimensional representation) learned from the data. They labeled this new feature set *score representation* and when using this new feature set for classification, they improved the performance of prior 3-class sentiment classification research (Figure A.4 (b)).

In this paper we follow a methodology similar to [3] in order to give structure to the input data. The reviews that the contributors have written for each object are the plain text format input. Using the aspect identification subsystem in



(a) Aspect identification

(b) Sentiment identification

Figure A.4: Aspect-level framework for sentiment analysis per Farhadloo and Rolland [3]. Aspects are identified by clustering the sentences using BON representation. Sentiment are identified using score representation.

Figure A.4 (a), a number of aspects are extracted. These aspects, along with the sentiment identification subsystem in Figure A.4 (b), are used to give structure to the data. For each contributor, the frequency count of positive, neutral, and negative sentiments for each aspect in each sentence fragment is summarized across the entire contribution. For each aspect, the frequency counts of positive, neutral and negative sentiments are calculated using the existing 3-class sentiment classifier. Therefore, if K aspects have been extracted in the first stage of the aspect identification framework , each review will be transformed into a $3K$ vector using the described scheme. It is worth mentioning that the frequency counts of positive, neutral and negative sentiments for each aspect are normalized separately. One can view this transformation as creating a matrix from the free-form text review.

Highlights

- We present a method for modeling overall customer satisfaction from written opinions.
- Reviews are transformed to semi-structured data using sentiment analysis.
- Overall customer satisfaction is measured using a Bayesian approach based on reviews.
- Empirical experiments on data from TripAdvisor show an aspect accuracy of 88.3%.
- Average R^2 values for predicted overall customer satisfaction range from 0.892 to 0.999.