Estimating numerical scale ratings from text-based service reviews

Article *in* Journal of Service Management · June 2020

DOI: 10.1108/JOSM-06-2019-0167

CITATIONS
13

READS
403

4 authors:

Hsiu-Yuan Tsao
National Chung Hsing University
35 PUBLICATIONS **1,136** CITATIONS
SEE PROFILE

Ming-Yi Chen
National Chung Hsing University
17 PUBLICATIONS **526** CITATIONS
SEE PROFILE

Colin Campbell
University of San Diego
117 PUBLICATIONS **5,323** CITATIONS
SEE PROFILE

Sean Sands
Swinburne University of Technology
80 PUBLICATIONS **4,363** CITATIONS
SEE PROFILE

Estimating numerical scale ratings from text-based service reviews

Hsiu-Yuan Tsao and Ming-Yi Chen
National Chung Hsing University, Taichung, Taiwan
Colin Campbell
University of San Diego, San Diego, California, USA, and
Sean Sands
Swinburne University of Technology, Melbourne, Australia

Received 1 June 2019
Revised 16 November 2019
13 January 2020
19 January 2020
Accepted 22 January 2020

Abstract

Purpose – This paper develops a generalizable, machine-learning-based method for measuring established marketing constructs using passive analysis of consumer-generated textual data from service reviews. The method is demonstrated using topic and sentiment analysis along dimensions of an existing scale: lodging quality index (LQI).

Design/methodology/approach – The method induces numerical scale ratings from text-based data such as consumer reviews. This is accomplished by automatically developing a dictionary from words within a set of existing scale items, rather a more manual process. This dictionary is used to analyze textual consumer review data, inducing topic and sentiment along various dimensions. Data produced is equivalent with Likert scores.

Findings – Paired *t*-tests reveal that the text analysis technique the authors develop produces data that is equivalent to Likert data from the same individual. Results from the authors' second study apply the method to real-world consumer hotel reviews.

Practical implications – Results demonstrate a novel means of using natural language processing in a way to complement or replace traditional survey methods. The approach the authors outline unlocks the ability to rapidly and efficiently analyze text in terms of any existing scale without the need to first manually develop a dictionary.

Originality/value – The technique makes a methodological contribution by outlining a new means of generating scale-equivalent data from text alone. The method has the potential to both unlock entirely new sources of data and potentially change how service satisfaction is assessed and opens the door for analysis of text in terms of a wider range of constructs.

Keywords Services quality, Machine learning, Text mining, Sentiment analysis

Paper type Research paper

Introduction

Imagine you are the manager of a boutique hotel in midtown New York. You care deeply about improving the reputation of your hotel, so are keen to garner guest feedback. Unfortunately, you find that guests are reluctant to respond to surveys you send them after a stay, even if you try and incentivize them to do so. Some guests complain that surveys are too long and dull, so imagine if you could get the same data from simply asking guests to talk for a minute or two about their service experience. Alternatively, imagine being able to mine the feedback on review sites such as TripAdvisor.com and get the same quantitative data that surveys provide in order to easily see trends in different aspects of hotel performance. You know monitoring guest reviews is important, but reading through potentially hundreds of new online reviews posted about your hotel this month is time-consuming and difficult to quantify. As you sit manually reading through the dozens of new online reviews posted about your hotel in the last two days, you think: could guests' review experience be improved? And could there be a way to somehow glean the same data as a survey from the text buried in online reviews, perhaps using AI and machine learning?

We answer the questions posed by the hotel manager in the aforementioned vignette by developing a novel machine-learning-based method for rapidly inducing numerical scale

Journal of Service Management
Vol. 31 No. 2, 2020
pp. 187-202

1757-5818
DOI 10.1108/JOSM-06-2019-0167

ratings from text-based consumer service reviews from any source. This is particularly important given the need to rapidly assess the growing volume of service reviews in online environments and effectively respond to enhance the customer experience. The method we describe uses machine learning to generate numerical data from text that is equivalent to the scale ratings garnered from traditional Likert scales. Importantly, the method does so without the need for a manually created dictionary, instead using existing scale items to automatically generate and calibrate a dictionary using machine learning. This new method is valuable to both service-industry practitioners and academics given that the intangibility of services makes reviews more important than in the case of products (Berry and Parasuraman, 2004). To date, machine learning has been underused in addressing service experiences and assisting managers to adequately understand and respond to the customer experience. Our method enables service providers to entirely assess service quality through online reviews rather than surveys. This not only eases the burden on consumers who are only asked to complete one postexperience step, but also enables service providers to solely focus on generating reviews, which are known to have a host of positive effects on potential consumers (Blazevic *et al.*, 2013; Keiningham *et al.*, 2018). For academics, this method not only has the potential to unlock access to more data from text, but also offers a more naturalistic method of collecting consumer service ratings compared to traditional scales. In sum, this paper illustrates how machine learning can be applicable to both service researcher and practitioners alike.

In this paper, we first overview the importance and value online reviews offer. We then describe how service quality is defined and traditionally measured, before reviewing emerging methods of text analysis and the unique contribution our approach offers. In a first study we detail how our method operates, using both Likert scale and textual reviews from the same consumer to validate our method. We then apply our same approach to reviews from a set of real reviews of New York area hotels. Finally, we close by discussing the practical, methodological and potential theoretical contributions of our paper.

Service reviews are increasingly important

The emergence of user-generated content, blogs and other platforms allows free expression of ideas, the creation of content and broad sharing of information (Yoo *et al.*, 2018). In the current information era, a potential consumer will often peruse online consumer product reviews before deciding to buy a product or service. Online reviews exert influence on customers in many ways: such as a willingness to pay (Pavlou and Dimoka, 2006), product trust and loyalty (Awad and Ragowsky, 2008) and intentions to spread word-of-mouth (Klaus and Maklan, 2012). Li and Hitt (2008) point out that to many consumers' reviews are viewed as true, credible, useful and influential. Since customers regard online reviews as one of the most trustworthy sources of information (Filieri, 2015), they are an important tool for making purchase decisions. Consumer reviews alter the consumer decision-making journey and are often a factor affecting purchase intention. Reviews influence purchase decisions and may even affect corporate financial performance and stock prices (Tirunillai and Tellis, 2012). This demonstrated effect is resulting in managers placing more emphasis on consumer reviews on these platforms since they are strategically important to a firm's success.

In the hotel industry, an increasing number of studies are being conducted to gain insight into customer satisfaction through online reviews (Lu and Stepchenkova, 2012). Since online reviews can be featured in promotions and drive online sales, as well as are important to marketing strategy and reputation management, they play a vital role in the lodging industry (Schuckert *et al.*, 2015). In the past, survey cards in guest hotel rooms were used to collect consumer opinions on service quality. Currently, use of sites such as TripAdvisor has resulted in potentially hundreds or even thousands of reviews of an establishment being posted each month. While reviews are valuable to a firm's

reputation and ability to attract new customers (Balahur *et al.*, 2012), the sheer volume of user-generated content has grown very rapidly in recent years (Kaplan and Haenlein, 2010; Munzel and Kunz, 2014). Gleaning valuable information from such customer commentary becomes challenging when those hundreds or thousands of reviews must be manually read and interpreted each month.

One approach to the volume of data is to ignore the text of reviews and focus on available summary metrics such as a star rating or the star ratings sometimes provided on additional subdimensions. While this greatly simplifies analysis, such ratings are shown to greatly shortchange the information buried in the accompanying text review of a service provider (Racherla *et al.*, 2013). They also likely do not provide information on all the specific product characteristics favored or unfavored by consumers or any service-delivery information. For instance, a star rating alone may not give information on all aspects customers are satisfied with (i.e. facility, attitude of service personnel, etc.), and overall star ratings are known to have minimal correlation with a service encounter (Racherla *et al.*, 2013). In some cases, reviews may also not include star ratings or ratings along relevant subdimensions. Compared to customer ratings, text reviews provide more information about hotel accommodation experiences and customer perceptions. More in-depth information can be gained about satisfaction and dissatisfaction by careful analysis of long-form written text (c.f. Jiang *et al.*, 2013). The volume of reviews enabled by the Internet and user-generated context is giving rise to more efficient and cost-effective analysis approaches (Berezina *et al.*, 2012). We overview these in more detail next.

The current state of automated text analysis tools

Machine analysis of text can follow either an exploratory or a quantitative approach. Exploratory approaches such as word clouds and clustering (e.g. Lim and Maglio, 2018; Timoshenko and Hauser, 2019) can provide organic description of text but are not easily compared across time or particular cases since the structure of each analysis is different. This is because the results of such analyses depend on the very text that they analyze, making results unique and incomparable.

In contrast to exploratory tools, quantitative approaches to analyzing text can provide numerical results. This includes use of techniques to analyze the frequency of specific words appearing as well as the positivity or negativity of comments (Stringam and Gerdes, 2010). This facilitates comparison of numerical scores across time and cases. Despite this, barriers to their use still remain. First is the expense and effort required to manually develop, test and validate the dictionary of words that form the backbone of such approaches. The time and effort of creating such dictionaries limit how many are created and actively used. For instance, dictionaries used in quantitative text analysis are often limited to more general topics such as those sentiment (also referred to as polarity), emotions (Neviarouskaya *et al.*, 2011) and personality (e.g. IBM Watson's automated analysis of the big five personality traits). Boutique or niche dictionaries are much less commonly used, likely due to the expense and time involved in setting them up. A second barrier to using existing quantitative approaches to text mining is that resulting metrics are difficult to contextualize relative to other market research data. In other words, it is currently challenging to compare text analysis results against scores obtained through a typical Likert-style survey. This creates challenges if data on the same topic or construct are collected using different methods since they are unable to be directly compared.

Despite advances in both exploratory and more quantitative text analysis tools, they remain distinct and thus limited in their abilities (Ordenes and Zhang, 2019). Exploratory approaches are valuable for getting the "lay of the land" from a piece of text, but their unstructured nature makes analyses difficult to compare. In contrast, quantitative approaches such as sentiment analysis can provide an overall metric but do not provide

insight into performance on specific dimensions unless a custom dictionary is created. Before describing a method to overcome these limitations of existing text analysis techniques by combining them into an automated AI-driven method, we next overview how service quality is traditionally measured.

How service quality is traditionally assessed

At an abstract level, service quality refers to the degree of difference between consumer expectations and actual service received (Parasuraman *et al.*, 1985). In short, it is a measure of whether the standard of services provided met consumer expectations or not. Service quality includes both intangible and tangible dimensions, which are inseparable from consumption. This means that the purchase of services often lacks concrete, tangible measurement indicators (Helkkula *et al.*, 2012). For this reason, the standard approach to determining service quality level is a satisfaction measurement. Oliver (1981) points out that satisfaction is a mental state that combines expectations and a consumer's previous consumption experiences.

Service satisfaction falls into a type of data that is difficult to quantify, as service satisfaction can be measured from many different aspects (Helkkula, 2011). In past research, measurement dimensions include: the attitudes of service personnel and physical facilities, which often change with consumer needs and interests, as well as time. This leads to an increase in the number of items measured to include components such as entertainment, technology or universality. In the hotel industry, reviews on online platforms show that perceived quality is actually more important than the actual quality experienced by customers (Bradley *et al.*, 2015).

There are many kinds of scales for measuring service quality such as SERVQUAL and SERVPERF (Gilmore and McMullan, 2009) or EXQ (Klaus and Maklan, 2012), with SERVQUAL being commonly employed (Ladhari, 2009). The original SERVQUAL scale had ten dimensions measured across 97 items, but these were later reduced to five dimensions (tangibility, reliability, responsiveness, assurance and empathy) measured across 22 items. Researchers point out that when applying the SERVQUAL scale for combined industries, the number of scale items may need to be decreased or increased (Pitt *et al.*, 1995; Van Dyke, Kappelman; Prybutok, 1997). The lodging quality index (LQI) is a lodging-specific variant of SERVQUAL developed by Getty and Getty (2003). The reliability and validity of the LQI scale were verified, and the relative weights of the five dimensions in terms of overall service quality, satisfaction and behavioral intention were all confirmed. LQI is known to relate to customer satisfaction, loyalty, repeat purchases, favorable word-of-mouth recommendations and ultimately higher profitability. The LQI scale is assessed using traditional Likert-style response scales.

The aim of this paper is to describe a method of *inducing* LQI scale scores from consumer's written reviews alone. The approach we describe combines the power of sentiment analysis with the specificity provided by analyzing the frequency of specific words, all in an automated fashion. Most importantly, a traditional scale is used to rapidly develop and tune a dictionary using machine learning. Specifically, items from an existing scale are used as a starting point for dictionary generation and training data from the same scale is also then used to calibrate text analysis results. This enables much greater insight to be gleaned and structured from text analysis of reviews. Our approach responds to growing interest in utilizing customer-generated data to gain insights into research problems that have not been feasible through conventional methods. While we describe our machine-learning-based method in the context of measuring LQI, it is generalizable and could be used to assess other constructs as well. In our first study we describe this method, as well as validate it using numerical scale ratings and review text from the same consumers.

Study 1: method validation using single-source text and Likert data

For the purpose of developing a method to measure the dimensions of LQI from text reviews, we utilized single-source data comprised of free-form written reviews and scale-based ratings from the same participants. This data was collected through an online survey distributed using an online panel of customers ($N = 551$) who stayed at a hotel in New York City. Participants wrote a 600-word free-form written review as well as responded to all 20 items of the LQI scale (Getty and Getty, 2003) using seven-point Likert scales. The order of these two tasks was counterbalanced. At the end of the survey, participants were asked to provide demographic information including gender, age, ethnicity, education, marital status and annual household income. Upon completion, participants were debriefed and thanked. Descriptive statistics on the sample are listed in Table 1.

We conducted factor analysis to reduce the original five dimensions of LQI to four dimensions, based on a KMO (Kaiser–Meyer–Olkin) of 0.95 and a cumulative proportion of variance of 68%. The Cronbach α of each factor satisfied reliability (all αs > 0.70) as well. We named the four dimensions tangibility, reliability, assurance and empathy. For the rotated component matrix, please refer to Appendix 1. While we reduced the factors to four for parsimony following similar reductions of SERVQUAL (Pitt *et al.*, 1995; Van Dyke, Kappelman; Prybutok, 1997), it is important to note the method we detail next works equally well for any number of dimensions. For clarity, we use "LQI" to refer to numerical scores calculated from Likert-scale scores and "LQI-Induced" to refer to numerical scores induced from text analysis using our method.

Automated dictionary generation from existing scale items

The first step in the LQI-induced method involves generating a dictionary for each of the four dimensions of the LQI construct. Words from each LQI dimension's existing scale items were used as stems from which to automatically build dictionaries using semantically similar thesaurus words. Dictionaries, especially thesauruses, are commonly used to suggest additional related words when developing dictionaries for content analysis. However, solely relying on a thesaurus to develop synonyms of dimensions can be challenging since some word pairs are closer in meaning than others. Fortunately, *Roget's Thesaurus* provides structure to the synonyms that it lists. *Roget's Thesaurus* is composed of six primary classes, conceptualized as a *tree* containing over a thousand branches of individual "meaning clusters" or semantically linked words. Although words depicted are not strictly synonyms, they can be viewed as colors or connotations of a meaning or as a spectrum of a concept. Based on *Roget's Thesaurus*' underlying tree structure, a semantic similarity value can be calculated for any pair of words. The value of the semantic similarity is coded as follows: a

	N	%
Gender		
Male	258	46.8
Female	293	53.2
Age		
18–25	69	12.52
26–35	312	56.62
36–45	91	16.52
46–55	56	10.16
56–65	18	3.27
66–75	5	0.91

Table 1. Descriptive statistics of our sample

score of 16 indicates a high similarity between two words, a score of 12–14 indicates an intermediate similarity and a score below 10 indicates a low similarity (Jarmasz and Szpakowicz, 2003).

The foundation for the LQI-induced dictionaries we developed using semantic similarity scores is a set of seed words. In the case of marketing constructs, existing scale items and conceptual definitions were used to select seed words. For instance, seed words were identified from definitions of each dimension of the LQI scale. *Roget's Thesaurus* was then used to generate words that were conceptually similar to each dimension's seed words. Specifically, we retained those words with a semantic similarity value of either 14 or 16. This process resulted in the development of a final list of 643 words distributed across four dimensions. The complete list of words used in the LQI-induced method are shown in Appendix 2.

Categorization and analysis of review text

Text sentiment analysis measures someone's words to find out how they feel and is in some cases considered more revealing than surveys because it is a more organic analytical method (Pang and Lee, 2004). The performance of such sentiment classifiers is dependent on the domain or topic being analyzed (Gunter *et al.*, 2014). We developed a custom program in *R*-programming language to scan all of the collected textual data and compare it against the semantic similarity-based LQI-induced dictionary we developed earlier. Based on keywords in the dictionaries, the program identifies relevant sentences and assigns each to a dimension of the LQI construct. Next, text data classified into dimensions of each construct is analyzed using sentiment analysis. We employed the publicly available AFINN Sentiment Word List. This is a well-known list of English words manually developed by Finn Årup Nielsen, a researcher at the University of Denmark, over a two-year period (Nielsen, 2011). Specifically, the AFINN word list was used to rate the valence of each sentence using an integer between –5 and +5 based on word strength. Our automated system also identifies and reverses the sentiment score of sentences containing negative modifiers [1].

For example, the sentences shown in Table 2 were written by a participant. The keywords "desk" and "attention" in those sentences can be found in the "tangibility" and "empathy" dimensions of LQI-INDUCED, respectively. Further, the emotional words, in this case, "nice" +3 and "great" +3, in those sentences are rated from AFINN.

When a customer writes a review online, it may capture only one or two dimensions of a larger construct, resulting in missing data on other dimensions. While a traditional questionnaire typically captures data covering all of a construct's dimensions, missing data is still sometimes an issue in this methodology as well. We address the problem posed by missing data in text analysis by imputing data for a dimension using other participants' responses. Specifically, average values from other respondents were used to replace missing

Table 2.
Sample sentences illustrating how text analysis works

ID	Dimensions	Featured word	Featured sentence	Emotional word	Sentiment score
58	Tangibility	Desk	I must say that the front desk guys were extremely nice and helped whenever needed	Nice	+3
60	Empathy	Attention	Great hotel to stay in, beautiful, great attention to detail and the staff were helpful. Loved my stay	Great	+3

values. This mirrors the approach typically used for missing values in survey data (Tsikriktsis, 2005; Vriens and Melton, 2002).

Machine-learning calibration and validation

Next, the set of data collected from our online study was used as initial data for machine-learning calibration. This data includes paired free-form written text and Likert-scale response data, both assessing the same service construct. Following a machine-learning approach, this initial data was randomly divided into training and holdout samples. Text data from the training data set was then analyzed against the dictionary developed for each of the dimensions of the LQI-induced marketing construct. Phrases that match each dimension were compiled and were evaluated for sentiment. This information was then used to calculate LQI-induced raw scores for each of the evaluated dimensions.

The LQI-induced raw scores reflect the initial combing of the reviews against the dictionaries and require two more steps to become equivalent to LQI ratings from actual Likert-scale responses. First, the LQI-induced raw scores need to be converted to be on the same seven-point scale as the actual Likert-scale response. This is done through a simple mathematical transformation [2]. Next, even though the scores are now on the correct scale, they need to be calibrated using the Likert scores from the training data. This enables the algorithm to "learn" how the two data sources relate. To do this, the LQI-induced raw text scores were compared against the paired LQI Likert-rating data (third row of Table 3) in order to develop standard conversion values to equate LQI and LQI-induced scores. These conversion values are key to the process since they effectively enable the algorithm to "translate" between the two forms of data.

Finally, the entire approach was tested using a holdout sample. This involved running the holdout text data through the dictionary and then converting the raw text scores (first row of Table 3) using the conversion values derived from the first stage. The resulting estimated LQI-induced measurement scale scores (second row of Table 3) could then be tested against each participant's actual LQI measurement scale scores (third row of Table 3). If there was not a significant difference between the estimated LQI-induced and actual LQI measurement scale scores, then the process would be considered successful. As reported in the fourth row of Table 3, *t*-test results indicate no significant difference between the LQI-induced Likert-converted text scores and LQI ratings from actual Likert-scale responses. This means the technique could then be employed on text data alone going forward.

Study 2: further validation using TripAdvisor data

In order to further validate the method used in calculating LQI-induced scores, we analyzed data from real reviews of New York area hotels. Specifically, we sought to confirm that LQI-induced scores from reviews matched with real-world ratings. WebHarvy software was used to scrape information from TripAdvisor (stars of hotel, name of hotel, username, overall

Table 3. Result of paired *t*-tests comparing estimated measurement scale scores with actual measurement scale scores of LQI-induced dimensions

	Assurance	Empathy	Reliability	Tangibility
1. LQI-induced raw text scores	3.96	2.82	2.9	3.04
2. LQI-induced Likert-converted text scores	5.66	5.63	5.72	4.21
3. LQI rating from actual Likert-scale responses	5.76	5.71	5.83	5.80
4. *P*-value of paired *t*-test comparing LQI-induced Likert-converted text scores and LQI rating from actual Likert-scale responses	0.237	0.223	0.125	0.083

rating and comments) for 27 hotels in New York City covering hotels with star levels from 2 to 5. There were 123,793 reviews scraped from a list of hotels between January and February 2019 (see Appendix 3).

Analysis

Machine learning refers to statistics that are designed to predict a particular outcome. In our case this is whether guests are satisfied, neutral or dissatisfied based on an individual's overall rating being higher than 3, equal to 3 or below 3 on a five-point star rating, respectively. Machine learning operates by splitting a data set into training, holdout and validation components, with the training sample used to build and optimize a predictive model and the holdout sample used to evaluate the model's performance. This is done by using the model to predict outcomes in the holdout sample and then comparing those predictions to actual outcomes. Better performing models are those that more accurately predict actual outcomes.

Gradient boosting is a machine learning technique referring to a technique of iteratively combining weak learners (i.e. algorithms with weak predictive power) to form an algorithm with strong predictive power (Friedman, 2002). Gradient boosting is an algorithm that has been very successful in applied machine learning. This includes winning Kaggle competitions for structured or tabular data (Ben Taieb and Hyndman, 2014). In this validation study, we ran a repeated cross-validation repeat three times with fivefold resampling cross-validation using stochastic gradient boosting via the caret package in *R*.

Our dependent variable is a dummy response variable representing whether the individual overall rating for a hotel is classified as satisfied (class = 1), acceptable (class = 0) or dissatisfied (class = −1).

In our study, we focus on the impact of LQI-induced scores gleaned from consumers' TripAdvisor reviews (refer to Study 1 for details on the process used to calculate the scores for the assurance, empathy, reliability and tangibility dimensions). Table 4 provides a description of the variables.

Results

In order to demonstrate the effectiveness of our LQI-induced measure in predicting the overall rating of a hotel, we compare prediction accuracy across a range of review types. Specifically, we look across both positive and negative reviews, as well as by the star level of the different hotels in New York City (ranging from 2 to 5 stars) (see Table 5).

First, the overall prediction rate based on reviews with positive sentiment (0.85 and 0.89 for LQI-induced and TripAdvisor ratings, respectively) is much better than reviews with

Table 4. Description of variables used in machine learning validation

Variable name	Data type	Variable description
Stars	character	Level of stars for hotel (Ranging from 2-star to 5-star)
Assurance	numeric	Knowledge and courtesy of employees and their ability to convey trust and confidence (ranging from −5 to +5)
Empathy	numeric	Caring, individualized attention the firm provides its customers (ranging from −5 to +5)
Reliability	numeric	Services are consistent with those they claim to offer (ranging from −5 to +5)
Tangibility	numeric	Refers to general appearance and functionality of the tangible assets of the hotel industry (ranging from −5 to +5)
Rating	factor	Overall rating ranging from 1 to 5
Class	factor	1 = Satisfied, 0 = Accepted, −1 = Dissatisfied

negative sentiment (0.60 and 0.53 for LQI-induced and TripAdvisor respectively). Second, whether positive or negative, hotels with extreme levels of stars such as 2- or 5-stars exhibit better prediction rates on overall rating than 3- and 4-star hotels. Third, while the prediction ability of LQI-induced and an individual's TripAdvisor rating is almost equivalent for in the case of positive reviews, the ability of TripAdvisor to predict when sentiment is negative is poor (0.5 more or less). In contrast, LQI-induced is still able to predict when sentiment is negative, especially in the case of those hotels with 2-star ratings (0.80).

General discussion

Over time there has been a downward trend in survey response rates (Baruch and Holtom, 2008) and an increase in online reviews. As such, this research provides service operators and researchers with an alternative machine-learning-based approach to collect data from consumers. Specifically, we present a novel method for inducing numerical scale ratings from text-based consumer service reviews collected from any source. Such insight is valuable to both service-industry practitioners and academics. In particular, our method provides service operators with an alternative to traditional surveys as a means to assess service quality, from data collected through online reviews. The method enables aggregate data to be collected from any source (online reviews, social media) and is shown to be equivalent to that collected via surveys. This is particularly useful, as social media postings are also often plentiful and range from simple mention of products and brands to more subjective expressions of feelings or experiences that consumers might share. As AI and machine learning grow even more powerful, we expect that similar techniques will only grow in efficacy and value.

For service managers, our proposed method suggests that there is potential to use machine learning to more effectively mine consumer insights from textual data beyond mere sentiment. It is known that deeper methods of analyzing consumer opinion are needed that go beyond assessment of positive and negative binary sentiment (Berger and Milkman, 2012). However, ways of building appropriate tools and mechanisms are still being explored (Tsao *et al.*, 2019). The method we develop in this paper uses reviews to generate numerical data that is the same as the data provided by traditional Likert scales. The method we describe offers several benefits. For those working in the service industry, our method is a passive way to glean service quality metrics from the text in online reviews. This is important since it allows services to focus on generating reviews rather than devote customer – and firm – energy toward online surveys. Free-form reviews are also a more organic and natural form of response than traditional measurement scales. This is likely to improve response rates as well as data quality, outcomes that are important to both academics and practitioners alike. For academics, the method we outline also provides a means of unlocking structured numerical data from text – a data source that is often difficult to analyze in a detailed fashion.

From the theoretical perspective, machine learning offers two distinct ways of approaching a given problem. One type of machine learning, supervised learning, is a way to train machines through some pattern of input and the result, and then the machine learns patterns and makes predictions or judgments. A second form of machine learning, unsupervised learning, utilizes

Star level of hotels	Negative (LQI-induced)	Positive (LQI-induced)	Negative (TripAdvisor Rating)	Positive (TripAdvisor Rating)
2-star	0.80	0.93	0.54	0.93
3-star	0.52	0.83	0.56	0.90
4-star	0.48	0.74	0.53	0.80
5-star	0.62	0.88	0.50	0.92
Average	0.60	0.85	0.53	0.89

Table 5. Accuracy in predictive overall rating by hotel star level and review valence

machine learning algorithms to infer patterns from a data set without reference to known, or labeled, outcomes (Cambria *et al.*, 2017). The method we described in this study is a form of supervised learning, with inputs coming in the form of words in existing marketing scales and a thesaurus, which are used to construct dictionaries that are compared against textual data. Machine learning then assesses how this data compares with Likert score responses from the same consumer. The machine learning algorithm is able to learn how the two relate so that it can infer Likert score responses from textual data alone. In many ways this is a kind of genuine or real form of human intelligence rather than "artificial" intelligence.

On the basis of the result we obtained from our TripAdvisor data, the ability of LQI-induced ratings induced from textual data to predict consumers' overall rating of a significantly exceeds those based on binary sentiment alone. This is particularly true in the case of negative opinion. However, solely improving prediction is not typically the goal of marketing research. Instead, we often want to reveal deeper consumer opinions based on the different service aspects reflected in dimensions of LQI-induced. Fortunately, the stochastic gradient boosting algorithm that underpins our machine learning provides insight into the relative importance of all variables on the dependent variable of rating.

Although measurement scales such as SERVQUAL and LQI are widely applied in a variety of industry and cross-cultural contexts, there are many criticisms of such approaches. A major concern is the design of the questionnaire. For instance, SERVQUAL contains 22 expectation items and 22 perception items, which combine to make the scale a rather long 44 total items. Therefore, the passive technique we develop eliminates the problem of long questionnaire since consumers can simply write free-form responses. Even better is that in many cases consumer opinion can be induced through text mining of public reviews rather than formally administrated – and costlier – surveys. It is likely that reviews are a more natural and enjoyable way of sharing feedback than traditional surveys (Munzel and Kunz, 2014). Our validation of this approach using paired *t*-tests of the Likert data and data from analysis of the written textual data illustrates that the two data sources are numerically equivalent. This opens up a tremendous amount of potential applications for practitioners.

This research, like all, is subject to certain limitations. We illustrates our method using a large data set of hotels in New York City; however, further research might consider examining a variety of different types of hotels (e.g. metropolitan, rural, country), travel purpose (e.g. luxury, business, family travel) or even extend this work into different service setting (restaurants, tourist attractions). Some of the reviews we analyzed could have been written by bots. While such noise would only bias against the results we report, it is still important to acknowledge. In addition, while we base our analysis of the quality of accommodation and satisfaction on the LQI marketing scale, some marketers might want to revise the dimensions they explore to more closely examine what concerns them. We also encourage work to improve the sensitivity of the machine learning estimation we employ. While we apportioned the data into three levels, future work might attempt to estimate even more levels.

Notes

1. For instance, the two sentences "I'm so happy" and "'I'm not so happy" can be tested using AFINN sentiment at the following link: http://darenr.github.io/afinn/
2. Likert-converted text scores = ((LQI-induced raw score – (−5))/10)*7.

References

Awad, N.F. and Ragowsky, A. (2008), "Establishing trust in electronic commerce through online word of mouth: an examination across genders", *Journal of Management Information Systems*, Vol. 24 No. 4, pp. 101-121.

Balahur, A., Kabadjov, M., Steinberger, J., Steinberger, R. and Montoyo, A. (2012), "Challenges and solutions in the opinion summarization of user-generated content", *Journal of Intelligent Information Systems*, Vol. 39 No. 2, pp. 375-398.

Baruch, Y. and Holtom, B.C. (2008), "Survey response rate levels and trends in organizational research", *Human Relations*, Vol. 61 No. 8, pp. 1139-1160.

Ben Taieb, S. and Hyndman, R.J. (2014), "A gradient boosting approach to the Kaggle load fore casting competition", *International Journal of Forecast*, Vol. 30 No. 2, pp. 382-394.

Berezina, K., Cobanoglu, C., Miller, B.L. and Kwansa, F.A. (2012), "The impact of information security breach on hotel guest perception of service quality, satisfaction, revisit intentions and word-of-mouth", *International Journal of Contemporary Hospitality Management*, Vol. 24 No. 7, pp. 991-1010.

Berger, J. and Milkman, K.L. (2012), "What makes online content viral?", *Journal of Marketing Research*, Vol. 49 No. 2, pp. 192-205.

Berry, L.L. and Parasuraman, A. (2004), *Marketing Services: Competing through Quality*, Simon and Schuster, The Free Press, New York, NY.

Blazevic, V., Hammedi, W., Garnefeld, I., Rust, R.T., Keiningham, T., Andreassen, T.W., Donthu, N. and Carl, W. (2013), "Beyond traditional word-of-mouth: an expanded model of customer-driven influence", *Journal of Service Management*, Vol. 24 No. 3, pp. 294-313.

Bradley, G.L., Sparks, B.A. and Weber, K. (2015), "The stress of anonymous online reviews: a conceptual model and research agenda", *International Journal of Contemporary Hospitality Management*, Vol. 27 No. 5, pp. 739-755.

Cambria, E., Poria, S., Gelbukh, A. and Thelwall, M. (2017), "Sentiment analysis is a big suitcase", *IEEE Intelligent Systems*, Vol. 32 No. 6, pp. 74-80.

Filieri, R. (2015), "What makes online reviews helpful? A diagnosticity-adoption framework to explain informational and normative influences in e-WOM", *Journal of Business Research*, Vol. 68 No. 6, pp. 1261-1270.

Friedman, J.H. (2002), "Stochastic gradient boosting", *Computational Statistics and Data Analysis*, Vol. 38 No. 4, pp. 367-378.

Getty, J. and Getty, R.L. (2003), "Lodging quality index (LQI): assessing customers' perceptions of quality delivery", *International Journal of Contemporary Hospitality Management*, Vol. 15, pp. 94-104.

Gilmore, A. and McMullan, R. (2009), "Scales in services marketing research: a critique and way forward", *European Journal of Marketing*, Vol. 43 Nos 5/6, pp. 640-651.

Gunter, B., Koteyko, N. and Atanasova, D. (2014), "Sentiment analysis A market-relevant and reliable measure of public feeling?", *International Journal of Market Research*, Vol. 56 No. 2, pp. 231-247.

Helkkula, A., Kelleher, C. and Pihlström, M. (2012), "Characterizing value as an experience: implications for service researchers and managers", *Journal of Service Research*, Vol. 15 No. 1, pp. 59-75.

Helkkula, A. (2011), "Characterising the concept of service experience", *Journal of Service Management*, Vol. 22 No. 3, pp. 367-389.

Jarmasz, M. and Szpakowicz, S. (2003), "Roget's thesaurus and semantic similarity", *Proceedings of Conference on Recent Advances in Natural Language Processing (RANLP 2003), Borovets*, Bulgaria, September 2003, pp. 212-219.

Jiang, L., Yang, Z. and Jun, M. (2013), "Measuring consumer perceptions of online shopping convenience", *Journal of Service Management*, Vol. 24 No. 2, pp. 191-214.

Kaplan, A.M. and Haenlein, M. (2010), "Users of the world, unite! the challenges and opportunities of Social Media", *Business Horizons*, Vol. 53 No. 1, pp. 59-68.

Keiningham, T.L., Rust, R.T., Lariviere, B., Aksoy, L. and Williams, L. (2018), "A roadmap for driving customer word-of-mouth", *Journal of Service Management*, Vol. 29 No. 1, pp. 2-38.

Klaus, P. and Maklan, S. (2012), "EXQ: a multiple-item scale for assessing service experience", *Journal of Service Management*, Vol. 23 No. 1, pp. 5-33.

Ladhari, R. (2009), "A review of twenty years of SERVQUAL research", *International Journal of Quality and Service Sciences*, Vol. 1 No. 2, pp. 172-198, available at: https://doi.org/10.1108/17566690910971445.

Li, X. and Hitt, L.M. (2008), "Self-selection and information role of online product reviews", *Information Systems Research*, Vol. 19 No. 4, pp. 456-474.

Lim, C. and Maglio, P.P. (2018), "Data-driven understanding of smart service systems through text mining", *Service Science*, Vol. 10 No. 2, pp. 154-180.

Lu, W. and Stepchenkova, S. (2012), "Ecotourism experiences reported online: classification of satisfaction attributes", *Tourism Management*, Vol. 33 No. 3, pp. 702-712.

Munzel, A. and Kunz, W.H. (2014), "Creators, multipliers, and lurkers: who contributes and who benefits at online review sites", *Journal of Service Management*, Vol. 25 No. 1, pp. 49-74.

Neviarouskaya, A., Prendinger, H. and Ishizuka, M. (2011), "SentiFul: a lexicon for sentiment analysis", *IEEE Transactions on Affective Computing*, Vol. 2 No. 1, pp. 22-36.

Nielsen, F. (2011), "Afinn", available at: http://www2.imm.dtu.dk/pubdb/p.php?6010.

Oliver, R.L. (1981), "Measurement and evaluation of satisfaction processes in retail settings", *Journal of Retailing*, Vol. 57 No. 3, pp. 25-48.

Ordenes, F.V. and Zhang, S. (2019), "From words to pixels: text and image mining methods for service research", *Journal of Service Management*, Vol. 30 No. 5, pp. 593-620, available at: https://doi.org/10.1108/JOSM-08-2019-0254.

Pang, B. and Lee, L. (2004), "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts", *Paper Presented at the Proceedings of the 42nd annual meeting on Association for Computational Linguistics*.

Parasuraman, A., Zeithaml, V.A. and Berry, L.L. (1985), "A conceptual model of service quality and its implications for future research", *Journal of Marketing*, Vol. 49 No. 4, pp. 41-50.

Pavlou, P.A. and Dimoka, A. (2006), "The nature and role of feedback text comments in online marketplaces: implications for trust building, price premiums, and seller differentiation", *Information Systems Research*, Vol. 17 No. 4, pp. 392-414.

Pitt, L.F., Watson, R.T. and Bruce, K. (1995), "Quality: a measure of information systems effectiveness", *MIS Quarterly*, Vol. 19 No. 2, pp. 173-187.

Racherla, P., Connolly, D.J. and Christodoulidou, N. (2013), "What determines consumers' ratings of service providers? An exploratory study of online traveler reviews", *Journal of Hospitality Marketing and Management* Vol. 22 No. 2, pp. 135-161.

Schuckert, M., Liu, X. and Law, R. (2015), "Hospitality and tourism online reviews: recent trends and future directions", *Journal of Travel and Tourism Marketing*, Vol. 32 No. 5, pp. 608-621.

Stringam, B.B. and Gerdes, J. Jr (2010), "An analysis of word-of-mouse ratings and guest comments of online hotel distribution sites", *Journal of Hospitality Marketing and Management*, Vol. 19 No. 7, pp. 773-796.

Timoshenko, A. and Hauser, J.R. (2019), "Identifying customer needs from user-generated content", *Marketing Science*, Vol. 38 No. 1, pp. 1-20.

Tirunillai, S. and Tellis, G.J. (2012), "Does chatter really matter? Dynamics of user-generated content and stock performance", *Marketing Science*, Vol. 31 No. 2, pp. 198-215.

Tsao, H.Y., Chen, M.Y., Lin, H.C. and Ma, Y.C. (2019), "The asymmetric effect of review valence on numerical rating: a viewpoint from a sentiment analysis of users of TripAdvisor", *Online Information Review*, Vol. 43 No. 2, pp. 283-300.

Tsikriktsis, N. (2005), "A review of techniques for treating missing data in OM survey research", *Journal of Operations Management* Vol. 24 No. 1, pp. 53-62.

Van Dyke, T.P., Kappelman, L.A. and Prybutok, V.R. (1997), "Measuring information systems service quality: concerns on the use of the SERVQUAL questionnaire", *MIS Quarterly*, pp. 195-208.

Vriens, M. and Melton, E. (2002), "Managing missing data", *Marketing Research*, Vol. 14 No. 3, p. 12.

Yoo, S., Song, J. and Jeong, O. (2018), "Social media contents-based sentiment analysis and prediction system", *Expert Systems with Applications*, Vol. 105, pp. 102-111.

Appendix 1

	Dimension			
	1	2	3	4
T8	0.717	0.395	0.178	0.219
T5	0.709	0.180	0.307	0.258
T4	0.687	0.293	0.327	0.208
C6	0.676	0.294	0.164	0.396
T7	0.667	0.235	0.273	0.228
T6	0.540	0.431	0.220	0.224
R2	0.228	0.705	0.251	0.191
T2	0.331	0.701	0.210	0.157
C3	0.204	0.676	0.337	0.296
C2	0.307	0.615	0.342	0.303
C5	0.386	0.545	0.266	0.295
RES3	0.380	0.524	0.319	0.318
R1	0.158	0.232	0.799	0.180
T1	0.394	0.128	0.756	0.174
C1	0.209	0.225	0.691	0.340
RES1	0.168	0.424	0.681	0.263
COMM4	0.113	0.065	0.260	0.867
COMM3	0.289	0.217	0.171	0.803
COMM2	0.345	0.246	0.128	0.577
COMM1	0.055	0.344	0.443	0.552

Table A1. Rotated component matrix

Appendix 2

Tangibility		Reliability	Assurance	Empathy
Accordance	Integration	Accuracy	Abutment	Accord
Acuteness	Junction	Adherence	Affiance	Acknowledgment
Amount	Justice	Allegiance	Affidavit	Acumen
Analysis	Kiss	Assurance	Affirmation	Affection
Animalism	Lick	Attachment	Agency	Affinity
Apotheosis	Manifestation	Authenticity	Allegation	Alliance
Appearance	Manipulation	Candor	Assertion	Appreciation
Applicability	Mark	Collateral	Asseveration	Attraction
Application	Material	Compact	Attestation	Benevolence
Aptness	Materiality	Confidence	Averment	Clemency
Archetype	Materialness	Conscientiousness	Avouchment	Click
Attention	Matter	Constancy	Avowal	Closeness
Avatar	Meaning	Contract	Back	Comfort
Bearing	Moderation	Correctness	Bargain	Commiseration
Being	Nicety	Covenant	Base	Communion
Blow	Nudge	Cover	Bed	Compassion
Body	Object	Credibility	Bedding	Compatibility
Brush	Objectiveness	Custody	Betrothal	Comprehension
Caress	Organization	Defense	Block	Compunction
Carnality	Palpability	Definiteness	Brace	Concord
Cast	Palpation	Devotedness	Buttress	Condolence
Characteristic	Particularity	Devotion	Certificate	Congeniality
Charitableness	Pat	Durability	Charter	Connection
Charity	Peck	Duty	Chat	Consideration
Civility	Peculiarity	Earnestness	Chitchat	Correspondence
Clearness	Percussion	Efficiency	Collar	Cotton
Clot	Personification	Equity	Colloquy	Dejection
Collection	Pertinence	Exactitude	Column	Discernment
Collision	Petting	Exactness	Confabulation	Distress
Compactness	Phenomenon	Fairness	Confirmation	Drift
Concentration	Propriety	Faith	Consent	Enthusiasm
Conformation	Prosopopoeia	Faithfulness	Consultation	Excitement
Congruence	Protoplasm	Fealty	Contention	Favor
Constituents	Push	Fidelity	Conversation	Fondness
Contact	Qualification	Frankness	Declaration	Forbearance
Contingence	Quality	Heartiness	Deposit	Grace
Contrast	Quantity	Homage	Device	Gratitude
Corporeity	Quintessence	Honesty	Discussion	Harmony
Courtesy	Rationality	Honor	Engagement	Heart
Crash	Reasonableness	Immunity	Footing	Humanity
Decency	Refinement	Impregnability	Foundation	Insight
Decorum	Relativity	Incorruptibility	Fulcrum	Intuition
Demonstration	Relevancy	Insurance	Gage	Judgment
Density	Revelation	Integrity	Groundwork	Kindliness
Detachment	Rigor	Invulnerability	Guaranty	Kindness
Diagnosis	Rub	Legitimacy	Guide	Leaning
Difference	Rubbing	Loyalty	Hold	Lenity
Differential	Scratch	Mastery	Lining	Link
Disclosure	Seemliness	Morality	Lock	Melancholy
Discrepancy	Sensuality	Obedience	Maintenance	Mercy
Discretion	Separation	Openness	Marriage	Obligation

(*continued*)

Table A2.
Dictionary Words in LQI-Induced Dimensions

Tangibility		Reliability	Assurance	Empathy
Discrimination	Sharpness	Patriotism	Means	Observation
Disinterest	Shock	Plainness	Medium	Partiality
Disinterestedness	Show	Preciseness	Mouthful	Passion
Dispassion	Sign	Principle	Oath	Penetration
Display	Significance	Probity	Parole	Perception
Dissemblance	Stiffness	Promise	Pillar	Perspicacity
Dissimilarity	Stroke	Purity	Pipe	Philanthropy
Dissimilitude	Structure	Rectitude	Platform	Pity
Distinction	Stuff	Redemption	Pledge	Quarter
Divergence	Substance	Refuge	Plight	Rapport
Divergency	Substantiality	Resolution	Pole	Recognition
Division	Substantialness	Responsibility	Post	Relationship
Due	Sum	Retreat	Precaution	Rue
Earmark	Symbol	Right	Predication	Ruth
Element	Symptom	Safeguard	Preservation	Sadness
Embodiment	Tact	Safekeeping	Profession	Sagacity
Embrace	Tactility	Safety	Promissory note	Sapience
Encompassment	Taction	Salvation	Pronouncement	Softness
Entity	Tap	Sanctuary	Prop	Solace
Epitome	Taste	Scrupulousness	Rain or shine	Sorrow
Equitableness	Thing	Security	Rampart	Soul
Estimation	Tolerance	Shelter	Ratification	Sympathy
Example	Touching	Shield	Recognizance	Tenderness
Exemplar	Type	Sincerity	Reinforcement	Testimonial
Exemplification	Unlikeness	Skill	Report	Thanks
Explanation		Skillfulness	Rest	Tribute
Exposure		Solidarity	Rib	Understanding
Expression		Solidity	Rod	Union
Feel		Soundness	Shoo	Unity
Feeling		Stability	Shore	Vision
		Straightness	Stake	Warmth
Fondling		Strength	Stanchion	Wisdom
Form		Strictness	Statement	Yearning
Formation		Subjection	Stave	Zeal
Graze		Submission	Stay	
Grope		Tie	Stipulation	
Heaviness		Token	Substratum	
Hedonism		Troth	Substructure	
Heed		Trustworthiness	Sustentation	
Hit		Truth	Swear	
Hug		Truthfulness	Swearing	
Impact		Uprightness	Talk	
Impartiality		Validity	Testament	
Importance		Veracity	Testimony	
Incarnation		Verity	Timber	
Inclusion		Virtue	Toast	
Incorporation		Ward	Undertaking	
Indication		Warrant	Vindication	
Individual		Weight	Vow	
Individuality		Wit	Warranty	
Instance			Word	

Table A2.

Appendix 3

Table A3. Complete List of Hotels Scraped and Number of Reviews

Star rating	Hotel	# Reviews
5	The Towers at Lotte New York Palace	1737
5	Crosby Street Hotel	859
5	Refinery Hotel	6661
5	The Knickerbocker Hotel	4591
5	The Michelangelo Hotel	4205
5	InterContinental New York Times Square	4,981
4	The Broom	584
4	The Sherry-Netherland Hotel	793
4	YOTEL New York	12,394
4	Crowne Plaza Times Square Manhattan	11,181
4	New York Marriott Marquis	12,526
4	Waldorf Astoria New York	11,590
4	The Roosevelt Hotel	10,885
4	Grand Hyatt New York	10,640
3	Hotel Giraffe by Library Hotel Collection	4,604
3	Residence Inn New York Manhattan/Central Park	1,103
3	Residence Inn New York Manhattan/Times Square	2055
3	Fairfield Inn and Suites New York Manhattan/Times Square	1,421
3	Washington Square Hotel	1443
3	Holiday Inn NYC–Manhattan 6th Avenue–Chelsea	1,479
3	Amsterdam Court Hotel	3840
2	Casablanca Hotel by Library Hotel Collection	6,554
2	Hotel 50 Bowery NY	1,193
2	Candlewood Suites New York City Times Square	1,390
2	Fifty Hotel and Suites	1,542
2	Pod 39 Hotel	2652
2	The Hotel @ New York City	890
Total		123,793

Corresponding author
Colin Campbell can be contacted at: colincampbell@sandiego.edu