Online Reviews as a Measure of Service Quality

Completed Research Paper

Biagio Palese
Louisiana State University
Baton Rouge, LA, United States
bpales1@lsu.edu

Gabriele Piccoli
Louisiana State University
Baton Rouge, LA, United States
gpiccoli@lsu.edu

Abstract

The proliferation of socialized data offers an unprecedented opportunity for designing customer service measurement systems. In this paper we address the problem of adequately measuring service quality using socialized data. The theoretical basis for the study is the widely used SERVQUAL model. The analysis is based on a database of online reviews generated on the website of the leading price comparison engine in Italy. We use a weakly supervised topic model to extract relevant dimensions of service quality from the user-generated content. Despite its exploratory nature the study offers two contributions. First, it demonstrated that socialized textual data, not just quantitative ratings, provide a wealth of customer service information that can be used to measure the quality offered by service providers. Second, it shows that the distribution of topics in opinions differs significantly between positive and negative reviews. Specifically, we find that concerns about merchant responsiveness dominate negative reviews.

Keywords: Online review, Service quality, SERVQUAL, Text mining, Topic model

Introduction

Since its commercialization in 1993, the Internet has dramatically changed people's behavior. Today we communicate by instant messaging, sharing pictures on social networks, and "tagging" our geolocation. More fundamentally, the Internet has altered how people make decisions. The emergence of the smartphone ecosystem and widespread connectivity has also changed the manner in which we procure goods and services. At the same time, the variety of products and services available to customers via the online channel is constantly increasing (Xu et al., 2013).

Brick and mortar organizations are forced to move online in order to prevent the loss of market shares. However, their lack of technical knowledge and experience with operating online combined with the different nature of online transactions can make this transition problematic, especially when it comes to service quality.

Customer service remains a key determinant of e-commerce success (DeLone and McLean, 2004; Xu et al., 2013) and drives customer satisfaction in online transactions (Cenfetelli et al., 2008, Xu et al., 2013). Service quality measurement has always been critical for organizations, but it has been historically limited by difficulties in collecting customers' opinions. However, with the rise of user generated content over the last decade, as well as the immediacy with which online customers can socialize their opinions on providers' websites, online review platforms and social media enable new approaches to service quality measurement. Socialized data is information that individuals knowingly and willingly share. Online reviews are a common form of socialized data, representing spontaneously shared opinions by customers on review platforms (Mudambi and Schuff, 2010).

To date, much of the literature on online reviews has focused on how they affect customer decisions. Much less work has examined how reviews can be used as a form of intelligence for gathering information for an organization. This is due to the fact that it is difficult to extract useful knowledge from large amounts of information (McAfee and Brynjolfsson, 2012). An effective measurement of service quality must be based on customers' experiences (Petter et al., 2012), thus making the content of online reviews particularly suitable.

Our work focuses on the textual elements of online reviews as a customer service measurement mechanism and offers two contributions. First, we use topic modeling, an emerging text mining approach, to extract from online reviews latent thematic structures that appropriately measure service quality. Specifically, we demonstrate that unstructured textual review data can be organized along the five elements of the widely accepted SERVQUAL model (Parasuraman et al., 1988). Second, we show that the different SERVQUAL dimensions have unequal impact on overall service evaluation in online reviews. This finding adds nuance to previous work that focused on aggregate measures of service rather than the contribution of each service quality dimensions (Luo et al., 2012).

Theoretical Framework

Service quality

Quality assessment is an important cross-disciplinary area of research in information systems, marketing and operations management. Early work focused on the quality measurement of physical products and tangible goods. In the second half of 20th century researchers developed systems to measure the quality of services (Gronroos, 1984; Parasuraman et al., 1985) because they recognize their unique characteristics of intangibility, heterogeneity, and inseparability. The literature provides definitions of service quality. From one perspective, service quality comprises of technical quality – what the customer is actually receiving from the service – and functional quality – the manner in which the service is delivered (Gronroos, 1982). From another perspective, service is co-produced between a provider and the recipient along three dimensions (Lehtinen and Lehtinen, 1982): physical quality (physical aspects of the service), corporate quality (company's image or profile), and interactive quality (interaction between contact personnel and customers). Based on the difference between initial customer expectation and actual perception, the SERVQUAL model (Parasuraman et al., 1985) is a seminal contribution to service quality measurement. After multiple refinements SERVQUAL (Parasuraman et al., 1988) coalesced on five dimensions: reliability (the ability to perform the promised service dependably and accurately); responsiveness (the willingness to help customers and provide prompt service); tangibles (the physical facilities, equipment, and appearance of personnel); assurance (the knowledge and courtesy of employees and their ability to inspire trust and confidence); and empathy (the caring, individualized attention the firm provides its customers). Since the introduction of SERVQUAL, there has been substantial research focused on testing the model and developing scales that are able to reliably measure service quality (Ladhari, 2009). SERVQUAL has been validated in various industries and it remains the most used instrument to assess the quality of service for both researchers and practitioners (Ladhari, 2009). We are aware that the SERVQUAL model received not only ample consensus, but also some critics over the years. In particular, Cronin and Taylor (1992) developed the competing SERVPERF model to measure only customers' perception. In this paper, it is not our intention to enter in the debate on which model developed in literature is better. We note that SERVPERF and SERVQUAL are grounded in the same dimensions. Our focus is on using those same dimensions to investigate their relevance in the text of online reviews. One of our innovations is to extract the dimensions of service quality not from surveys, as it is traditionally done, but rather algorithmically from text that customers socialized voluntarily when sharing their review. So, given the exploratory nature of the study and the characteristics of the sample analyzed, we decided to choose the most widely investigated instrument available – namely SERVQUAL.

Online transactions uncertainty and new sources of information

The assurance of high service quality is a priority for companies that move to online commerce (Xu et al. 2013). Quality service is critical in e-commerce to increase channel usage (Devaraj et al., 2002), customer loyalty (Gefen, 2002), and customer satisfaction (Cenfetelli et al., 2008; Tan et al., 2013). Customer service is particularly crucial for small and medium enterprises with low visibility (Luo et al., 2012). Yet despite its importance, we have limited knowledge about the determinants of online customer service quality (Xu et al., 2013, Petter et al., 2013).

E-commerce transactions are computer mediated and the absence of physical interaction results in high uncertainty for customers. Offline physical transactions are personal and contact based (Xu et al. 2013), thereby providing a multitude of information cues to customers. Many of these cues are lacking in online transactions, historically leading to customer uncertainty that discourages e-commerce (Ba et al. 2003) and limits online trust (Gefen et al., 2008).

The above limitations are counterbalanced by website design (Jiang and Srinivasan, 2012) and the increasing availability of user generated data. First, the rise of Web 2.0, and later, the shift to the mobile platform, supported the emergence of online product review platforms (e.g. TripAdvisor, Yelp.com, Amazon etc.). These platforms offer consumers the opportunity to post product reviews with content in the form of numerical star ratings and open-ended, customer-authored comments (Mudambi and Shuff, 2010). The computer-mediation of customer service automatically generates data in a digital form (Piccoli and Watson, 2008). This data can potentially impact not only individual users' decision-making processes but also guide organizations' managers in making strategic decisions (Piccoli and Pigni, 2013).

While much of the academic research has focused on consumer use of online reviews and the impact they have on their decisions, online reviews are an important source of unfiltered customer intelligence. Until the emergence of user generated content, it was difficult and time consuming to gather the data needed to measure service quality with customer surveys being the primary instrument of data collection. However, customers are increasingly overwhelmed by company communications (e.g., email, phone calls, robo-calls) soliciting their opinion. Even when incentives are offered or remuneration is provided to respondents, customer service surveys are plagued by limitations such as low response rates, small samples, and high expense (Wright, 2005).

Socialized data, those data that individuals knowingly and willingly share, are by definition broadcasted via online media, thus containing information essential for companies, but also available to other entities (e.g., competitors, customers, suppliers). The IT-mediation of these contributions makes them different from traditional word of mouth. In fact, while traditional word of mouth occurs through deep information exchanges between a small number of individuals, online reviews engender difficulties in navigating among thousand of these contributions and heuristics, such as examining aggregate quantitative evaluations (i.e., average rating of a product) and the close examination of only a few commentaries (Ghose and Ipeirotis, 2006). Moreover, online reviews ratings distribution is bimodal, so the average ratings cannot be considered an accurate measure (Hu et al., 2006) and an overall neutral rating is not always representing a neutral opinion (Jabr and Zheng, 2014). At the same time, also reading just a few recent reviews is unlikely to yield accurate perception of a service quality. While this is a problem for customers seeking decision-making aids, it is even more problematic for organizations attempting to measure customers' perception of service quality. Extracting and summarizing the service-specific thematic structure hidden in online reviews textual elements provides a solution to this quandary. The first objective of our exploratory work is to *demonstrate whether the dimensions of the SERVQUAL model can be extracted directly from the textual component of the online reviews* using topic modeling. In order to assess the quality of our methodology, we perform a validation (Validation procedure 1, Appendix B) of our topic model results. Our second objective is to *analyze which SERVQUAL dimensions influence overall customer evaluation the most*. As discussed above, online transactions engender increased levels of customer uncertainty and limit trust. Currently there is no research that we are aware of that empirically demonstrates the relative importance of service dimensions on customer satisfaction. Consequently, we structure this question as an exploratory investigation.

Methodology

Research context

Our research is set in the context of a price comparison website. The company enables users to search for products and it returns a list of all merchants carrying it, along with price and customer reviews data (Figure 1). Customers who want to make a purchase are directed to the merchant's website to place an order, and the merchant fulfills the transaction directly. In this paper, we ignore the reliability issue generally associated with this type of contributions because only those customers with verified purchases can write a review assessing their experience with the merchant on the price comparison engine's own website. These reviews consist of an overall rating assessment as well as the following five dimensions: ease of contact with the merchant, ease of purchasing from the merchant, ease of merchant website navigation, product delivery speed and customer service. Customers can also provide commentary in a free form text field.

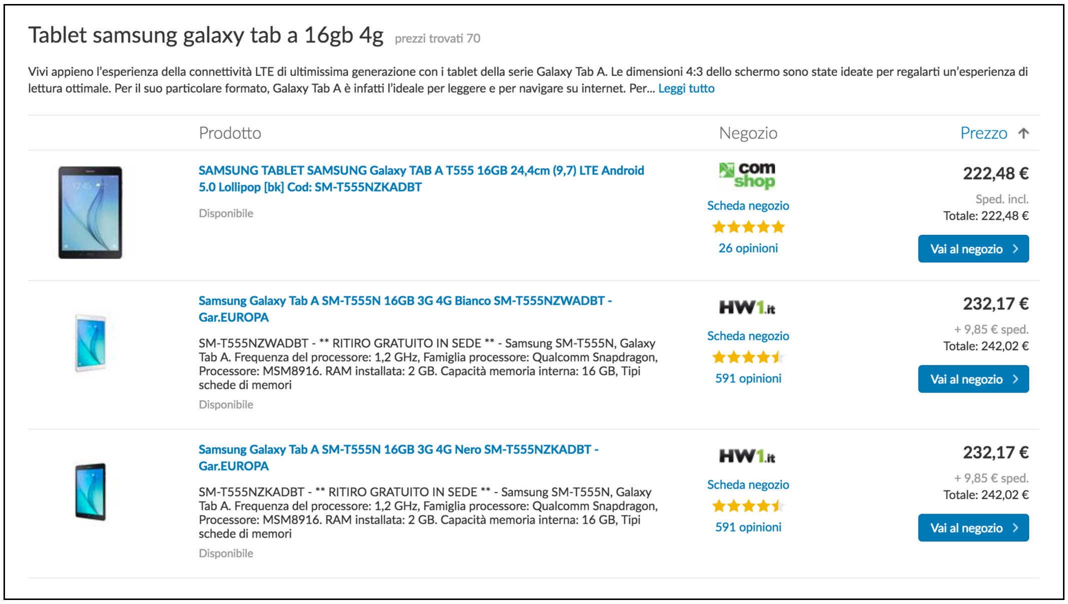

Figure 1: Search results page

Data analysis: Topic model

With few exceptions (Archack et al., 2011; Duan et al., 2013; Piccoli and Ott, 2014), previous research has taken a narrow methodological focus, analyzing the quantitative aspects of reviews and neglecting the rich data available in the review prose. However, machine learning researchers developed multiple systems that are able to automatically extract, evaluate, and present opinions in ways that are both helpful and interpretable. Early approaches to automatically extract and interpret review text have focused on determining either the overall polarity (i.e., positive or negative) or the sentiment rating (e.g., one-to-five stars) of a review. However, only considering coarse overall ratings fails to adequately represent the multiple dimensions of service quality on which a company can be reviewed. Topic modeling, a technique that extracts the hidden thematic structure from the documents, offers a solution (Blei, 2012).

Topic models are "[probabilistic] latent variable models of documents that exploit the correlations among the words and latent semantic themes" (Blei and Lafferty, 2007). Topic models can extract surprisingly interpretable and useful structures without any "understanding" of language by the computer. A document is modeled as a mixture of topics. This intuitive explanation of document generation is modeled as a stochastic process, which is then "reversed" (Blei and Lafferty, 2009) by machine learning techniques that return estimates of the latent variables. Given these estimates, it is possible to perform information retrieval or text mining tasks on the corpus. The interpretable topic distributions arise by computing the hidden structure that likely generated the observed collection of documents (Blei, 2012). In our analysis, we use a weakly supervised approach to topic modeling using Gibbs-sampling. Sampling-based algorithms attempt to collect samples from the posterior distribution to approximate it with an empirical distribution (Griffiths and Steyvers, 2004). In Gibbs sampling specifically, a Markov chain is constructed. This is a sequence of random variables, each dependent on the previous one, whose equilibrium distribution is the posterior (Steyvers and Griffiths, 2007).

Experimental setup: Dataset and Preprocessing

We obtained 74,775 online reviews provided from the leading Italian online price comparison company. The sample includes all of the reviews that the company had accumulated from its inception up to the moment we started our study, covering a period of 8 years. The reviews refer to different small online shops available for price comparison on the company website. The sizes of the companies in our sample make it even more relevant to them to provide a high quality service. The database presents a J distribution in which positive reviews (58,988) appear over ten times more frequently than the negative reviews (5,696). In this section, we consider negative reviews those with one-star rating, while positive reviews are those with five stars.

Online review contents are unstructured data, so it is necessary to apply standard preprocessing steps in order to perform text analysis on them. All analysis was performed using R. Through standard pre-processing, using the *tm* package (Feinerer and Hornik, 2015), we remove singleton words, stop words, numbers, and exclude reviews that were too short - less than 50 words (Lu et al., 2011), bringing the proportion of negative to positive reviews from 1/10 to 1/4. This confirms that when reviews are positive, their length is shorter on average (Piccoli and Ott, 2014). After removing non-Italian reviews using the *textcat* package (Hornik et al., 2013), we were left with 27,117 reviews. The dataset was then tokenized using the *MC_tokenizer* (Feinerer and Hornik, 2015) into unigram and was split into sentences using the *strsplit* function resulting in a total of 122,919 sentences.

Method: Multi-Aspect Sentence Labeling using weakly supervised topic models

The empirical approach used in this work is based on Lu et al. (2011). With a weakly supervised topic model, we performed a multi-aspect sentence labeling using the *topicmodels* packages (Gruen and Hornik, 2011). The first phase of multi-aspect sentiment analysis is usually aspect identification. In this paper we used the dimensions of SERVQUAL as aspects since we want to extract them from the reviews' content. This approach utilizes only minimal prior knowledge, in the form of seed words, to enforce a direct correspondence between topics and aspects. We selected words using only nouns associated with the essence of SERVQUAL dimension. We selected these terms directly from the vocabulary of our corpus. The seed words include only the most frequent and descriptive nouns. Eliminating adjectives reduced the risk of misinterpretation of the topics, since adjectives can relate to any of the SERVQUAL dimensions (Table A, Appendix A). To encourage the topic model to learn latent topics that correlate directly with aspects, we augmented them with a weak supervised signal in the form of aspect-specific seed words. We use the seed to define an asymmetric prior on the word-topic distributions. This approach guides the latent topic learning towards more coherent aspect-specific topics, while also allowing us to utilize large-scale unlabeled data. The prior knowledge (seed words) for the original LDA model is defined as a conjugate Dirichlet prior to the multinomial word-topic distributions β. By integrating with the symmetric smoothing prior η, we define a combined conjugate prior for each seed word w in $\beta \sim \text{Dir}(\{\eta + C\_w\}: w \in \text{Seed})$, where C_w can be interpreted as a prior sample size (i.e., the impact of the asymmetric prior is equivalent to adding C_w pseudo counts to the sufficient statistics of the topic to which w belongs). The pseudo count C_w for seed words was heuristically set to be 3000 (about 10% of the number of reviews following Lu et al., 2011). Assuming that the majority of sentences were aspect-related, we set the number of topics K to six, thereby allowing five topics to map to SERVQUAL dimensions and a residual unsupervised "background" topic. The six labels associated with each sentence are: reliability, responsiveness, tangibles, assurance, empathy and "background". In this work we sampled the models for 1,000 iterations, with a 500 iterations burn-in and a thinning of 10 iterations. We assigned the following value to topic model hyperparameters: $\alpha = 0.1$ and $\eta = 0.1$ (Lu et al., 2011).

We assumed that aspects are fixed following SERVQUAL dimensions and that each sentence of an online review typically addresses only one SERVQUAL dimension. Thus, we set a minimum threshold (0.6) to perform the classification, so the algorithm automatically labels each sentence with the most prevalent topic. Moreover, sentences that do not address any of the six topics above the threshold are considered "undefined". For example, we report a review from our sample with its translation:

"Acquisto andato a buon fine, sono davvero soddisfatto e felice di aver scelto questo sito! Imballo perfetto nulla da ridire. Prodotto arrivato in tre giorni come indicato sul sito, super affidabile!

Nonostante vivo in un piccolo paese del sud italia, per di più non ben collegato, e non in una grande città."

(Purchase went well, I am really satisfy and happy of choosing this website! Perfect packaging, nothing to complain. Product arrived in three days as indicated on the site, super dependable! Even if I live in a small village in the south of italy, in addition not well connected, and not in a big city.)

The above review has been classified as background (first sentence), tangibles (second sentence), reliability (third sentence) and "undefined" (fourth sentence).

The second objective of our work is to *identify the SERVQUAL dimensions that mostly influence the overall customer evaluation*. To do so, we computed new variables: review breadth, review depth, and review length measured as the number of words used in a review. In fact, the breadth is an indicator of the variety of topics utilized in each review. More precisely, breadth represents the number of different topics (from 0 to 6) discussed in each review by at least one sentence (Madlberger and Nakayama, 2013). However, in order to understand the impact of each topic we also examined review depth, defined as the number sentences used in each review, to describe the same topic (Madlberger and Nakayama, 2013). We then performed a multiple regression analysis to understand how these variables affect the online reviews' overall rating. In the next section we discuss our major findings.

Results

The output of topic modeling is a set of K topics predetermined by our weakly supervised approach. Each topic has a distribution for each term in our vocabulary. What characterized the topics is the terms distribution, as represented by the most frequent terms. The presence of the seeding terms and words related to them in the appropriate topic provides an indication of the efficacy of the seeding. However, this first indication is not sufficient to assess model validity. Formal validation was done by five independent raters (graduate students) unaware of the research objectives or the seeding process. The raters were given definitions of the five SERVQUAL dimensions and had to assign each unnamed topic represented by the ten most frequent words to the correct dimension (Validation procedure 1, Appendix B). The validation procedure results showed 93.3% accuracy in identifying the topics. In order to assess the reliability of the agreement between the respondents, we calculated Fleiss' kappa and showed that agreement is deemed almost perfect (Landis and Koch, 1977).

$$k \pm \text{NORMSINV}(1-\alpha)^{*}\text{s.e.}$$

$$\text{IC: } 0.858 \pm 0.095$$

Where k is the Fleiss' kappa, $\alpha = 0.05$ and s.e. is the standard error= 0.057

After demonstrating appropriate topic extraction from the reviews, we analyzed the number of sentences associated with each topic. At this point, we removed 30,742 "undefined" sentences that did not unambiguously represent one topic (i.e., no topic had a probability greater than 0.6). Thus, we found that responsiveness (20%) and empathy (22%) are the preponderant topics in our corpus. On the contrary, tangibles (12%) and assurance (13%) are discussed less often. The high accuracy of the validation and these results confirm that it is possible to extract coherent thematic structures from socialized data and that it is possible to extract customer perception of service along the dimensions of the SERVQUAL framework.

Our second research objective is to understand which of the dimensions of SERVQUAL had the strongest impact on overall customers' evaluations of the service quality provided by the merchants:

$$\begin{aligned} \boldsymbol{Rating} = {} & \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\ \boldsymbol{Review\ length} + \boldsymbol{\beta_2}\ \boldsymbol{Review\ breadth} + \boldsymbol{\beta_3}\ \boldsymbol{Reliability\ depth} \\ & + \boldsymbol{\beta_4}\ \boldsymbol{Responsiveness\ depth} + \boldsymbol{\beta_5}\ \boldsymbol{Tangibles\ depth} + \boldsymbol{\beta_6}\ \boldsymbol{Assurance\ depth} \\ & + \boldsymbol{\beta_7}\ \boldsymbol{Empathy\ depth} + \boldsymbol{\varepsilon} \end{aligned}$$

The results (Table 1) show that review length has a negative significant effect on overall review rating, while review breadth does not have a significant impact. Among topics' depth, only the depth of responsiveness and tangibles has a significant negative impact on the rating, while reliability, assurance, and empathy have a positive one. Looking at the estimates, it is clear the relevance of the responsiveness. In fact, if its depth increases by 1 then the overall rating will decrease by 0.65. The multicollinearity was tested using VIF. All the variables in the model have VIF smaller than 5 and the mean of the VIF is smaller than 2, indicating the absence of multicollinearity in our model.

Coefficients	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**	**VIF**
(Intercept)	4.612474	0.019369	238.134	< 2e-16***	
Review length	-0.004567	0.000170	-26.869	< 2e-16***	1.455675
Review breadth	-0.013899	0.011357	-1.224	0.221	3.059002
Reliability depth	0.229245	0.010702	21.420	< 2e-16***	1.518179
Responsiveness depth	-0.655666	0.007428	-88.275	< 2e-16***	1.658197
Tangibles depth	-0.079397	0.011340	-7.001	2.6e-12***	1.461323
Assurance depth	0.354146	0.011697	30.276	< 2e-16***	1.486494
Empathy depth	0.134628	0.009773	13.775	< 2e-16***	1.508776
(mean VIF)					1.735378

Significance levels: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 1: Multiple regression results

However, in order to better explain these findings, we examined the topic distributions. Overall, the reviews have a mean breadth of 3.486, indicating that users discuss between three or four different topics per review. Interestingly, responsiveness is discussed in the second lowest percentage among topics (Table 2). However, there are stark differences in depth by topic. Reviews discussing responsiveness are split about evenly between those with depth of 1 (53.76%) and those addressing responsiveness with more than one sentence. A quarter of reviews addressing responsiveness have depth greater than two (24.32%). Conversely, the other dimensions only have around 10% of reviews with more than two sentences dedicated to the same service quality dimension (reliability: 6.49%, tangibles: 7.96%, assurance: 4.14%, empathy: 11.81%). This result indicates that when customers discuss the responsiveness of the merchant, they emphasize this aspect of the service experience disproportionately more than any other topic.

	Depth by SERVQUAL dimension							Reviews	Dimension Proportion
	0	1	2	3	4	5	> 5		
reliability	15,441	8,496	2,422	566	157	27	8	11,676	43%
responsiveness	17,811	5,003	2,040	1,097	542	325	299	9,306	34%
tangibles	19,125	5,913	1,443	422	157	36	21	7,992	29%
assurance	17,738	7,121	1,870	330	48	9	1	9,379	35%
empathy	14,031	8,536	3,004	1,034	353	106	53	13,086	48%
background	16,054	8,602	2,020	343	78	13	7	11,063	41%

Table 2: Number of reviews divided by number of sentences associated to each topic.

It is the focus of negative reviews on responsiveness that explains the difference in distribution by topic (Figure 2). While negative reviews are only one fourth of the sample, they are dominated by sentences focusing poor service quality on the responsiveness dimension. Responsiveness is discussed in only 18.61% of positive reviews while 86.07% of negative reviews address it. This means that responsiveness is around seven times more frequent than assurance in negative reviews. Responsiveness is the only topic that presents a U shape (instead of the typical J distribution of review valence in our dataset). Moreover, in absolute terms, the negative peak is even higher than the positive one. The dominance of responsiveness in negative reviews suggests that, not only is it the most relevant topic, but also that it can dramatically affect rating distribution.

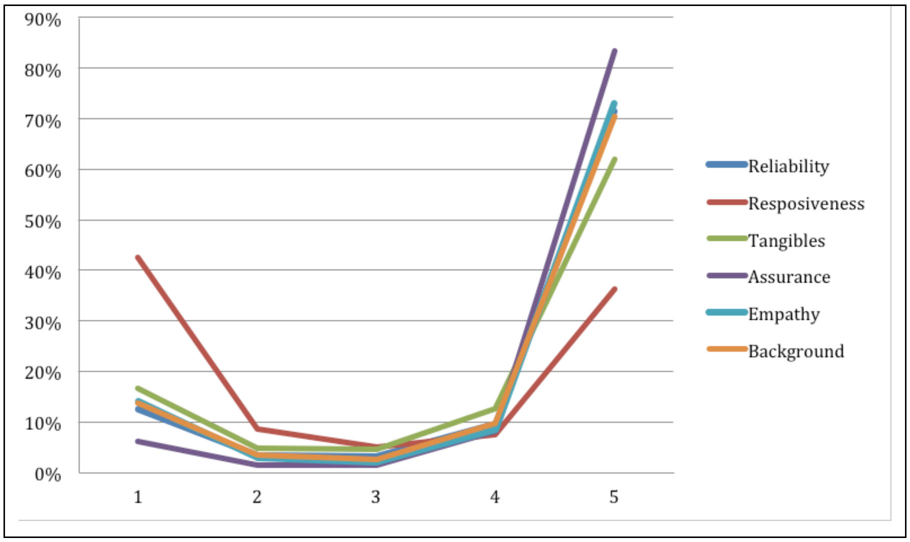

Figure 2: Topic distribution among reviews' rating

Practical Implications

These findings add nuance to previous studies that focused only on the aggregate measure of service quality (Luo et al., 2012) because they provide insights about each determinants impact on online reviews evaluation of service quality (Xu et al., 2013, Petter et al., 2013). In fact, the analysis showed that SERVQUAL dimensions have different distribution in terms of rating and depth. Companies looking to improve their service quality need to consider them to achieve their goal.

The above results have significant practical implications for the data providers, and by extension, for the design of online review systems. In a new validation procedure (Validation procedure 2, Appendix B), we asked to map SERVQUAL dimensions to the current evaluation system adopted by the company. The mapping, in this case, was performed by one world-renowned customer service expert and by five graduate students. The results of this validation (Table C, Appendix C) show that some of the system evaluation criteria used are too broad, while others are unable to capture any of the topics. Moreover, none of them accurately measures responsiveness: the most influential topic in our findings. The validation gives us an indication that the numeric system actually adopted by the company can provide misleading information about customer assessment of the overall service experience. We therefore propose a new evaluation system composed by questions that we have created on the basis of the results of our research. The mapping (Table D, Appendix C) in this case shows a higher accuracy in measuring the different topics, but also suggests that some changes are still necessary. However, the purpose here was only to show that our model is able to extract knowledge directly from customers' reviews and lead to service quality measurement systems that not only are theory-based, but also are more accurate.

Conclusion

Our exploratory study contributes to research on the use of the increasing wealth of digitally streamed data. Our results should also prove useful to designers and users of customer service systems. We believe that an organization that exploits social data spontaneously generated by their customers not only can improve service quality measurement, but also can have a better understanding of the aspects that influence their satisfaction expressed as an overall rating. In fact, the average of ratings, given their distribution in online reviews, can not be considered a reliable measure (Hu et al,. 2006) and even a neutral rating is not always representative of a neutral opinion (Jabr and Zheng, 2014). Moreover, in this way it will be possible to make decisions based on information gathered directly from their customers and avoid the current behavior of following what other companies do (Ostrom et al. 2015). An effective measurement of service quality must be based on customer experience (Petter et al., 2012). Furthermore, service quality evaluation systems should be able to map with reviews' topic content in order to improve customer experience and to increase measurement accuracy. Companies that want to achieve high customer service cannot ignore topics that effectively and heavily affect their evaluation. For example, the current evaluation system adopted by our data provider ignores responsiveness, the most influential topic for its users.

We also show that automated algorithms, like topic modeling, can be used to extract meaning from the huge amount of socialized data. In this way, we respond to the call to find applications of text mining capable of uncovering information not accessible with traditional methods (Ostrom et al., 2015). In fact, these new technologies enable the systematized assessment of service quality systems able to reliably measure all the aspects that influence customer evaluations.

Improvements in this direction can be beneficial for both the customers that generally make a decision based on the quantitative rating of inaccurate criteria, and to the organization gaining real time knowledge of customers' opinions. Furthermore, reducing the difficulties in navigating among those contributions (Ghose and Ipeirotis, 2006) can improve customer experience online.

In the future, researchers can focus on expanding upon our work by comparing multiple service quality models in order to assess the one that is more capable of accurately capturing the topics affecting customers' evaluation. Another extension would be to control other aspects that can impact the overall customer experience (price, product quality etc.). While our work uses Italian reviews, the language has no effect on the generalizability of our results. However, we plan to replicate this study using a database of English reviews and broaden the study to different industries.

Appendix A: Seeding and Topics

Seedwords

SERVQUAL dimensions	Seed words
Reliability	pacco (package), spedizione(shipment), consegna(delivery), ritardo(delay).
Responsiveness	mail, email, risposta(response).
Tangibles	sito(website), corriere(carrier).
Assurance	servizio(service), gentilezza(kindness), professionalità(professionalism)*, serietà(earnestness)*.
Empathy	cura(care), assistenza(assistance).

Table A: Seed words.

These words (*) have been used in the real seed words as professionalit and professionalita, seriet and serieta due to incorrectly identified encoding, non-ASCII letters were removed. English translation is reported in parenthesis.

Topics words

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
consegna* (delivery)	mail*	sito* (website)	servizio* (service)	acquisto (purchase)	prezzo (price)
spedizione* (shipment)	dopo (after)	corriere* (carrier)	seriet* (earnestness)	assistenza* (assistance)	acquistato (purchased)
ordine (order)	email*	prodotto (product)	professionalit* (professionalism)	negozio(shop)	euro
pacco* (package)	risposta* (response)	imballo (packaging)	molto (very)	cura*(care)	prodotto (product)
giorni (days)	ordine (order)	senza (without)	gentilezza* (kindness)	sito (website)	negozio (shop)
dopo (after)	giorni (days)	perfetto (perfect)	serieta* (earnestness)	prezzi (prices)	sito (website)
giorno (day)	stato (status)	stato (status)	professionalita* (professionalism)	prodotti (products)	samsung
stato (status)	ancora (yet)	arrivato (arrival)	ottimo (excellent)	acquisti (purchases)	acquisto (purchase)
ritardo* (delay)	prodotto (product)	pacco (package)	consegna (delivery)	consiglio (advice)	trovato (found)
arrivato (arrival)	disponibile (available)	problema (problem)	sempre (always)	dire (to say)	spedizione (shipment)

Table B: Topics with the ten most frequent words.

The words chosen for the seed are marked with *. English translation is reported in parenthesis.

APPENDIX B: Validations procedures

Validation Procedure 1

This validation procedure is composed on two distinct parts.

In the first we describe SERVQUAL, a methodology used to evaluate customer service.

In the second part we show you six topics described by the ten most frequent terms.

To validate the model, you need to assign the different topics to the five dimensions of customer service quality. Since there are six topics and only five dimensions, you need to name one topic.

Introduction to the Service Quality literature

In order to complete this procedure, it is very important to understand the SERVQUAL dimensions. Please read the definitions and examples below very carefully.

SERVQUAL Dimensions:

•**Reliability**: Ability to perform the promised service dependably and accurately.

Examples: 1) Providing services as promised. 2) Performing services right the first time. 3) Providing services at the promised time

•**Responsiveness**: Willingness to help customers and provide prompt and quick service.

Examples: 1) Readiness to respond to customers' requests. 2) Convenient business hours, easy to interact with the company. 3) Prompt service to customers. 4) Willingness to help customers.

•**Tangibles**: Appearance of physical facilities, equipment, personnel, and communication materials.

Examples: 1) Products are in perfect condition when delivered to the customer. 2) Packaging is visually appealing and in good conditions. 3) Employees who have a neat, professional appearance. 4) Visually appealing materials associated with the service.

•**Assurance**: Knowledge and courtesy of employees and their ability to inspire trust and confidence in customers.

Examples: 1) Employees who instill confidence in customers. 2) Making customers feel safe in their transactions. 3) Employees who are consistently courteous. 4) Employees who have the knowledge to answer customer questions.

•**Empathy**: Caring, individualized attention the firm provides its customers.

Examples: 1) Giving customers individual attention. 2) Employees who deal with customers in a caring fashion. 3) Having the customer's best interest at heart. 4) Employees who understand the needs of their customers.

Topics identification

Here you find six different topics, please assign them to a dimension and write it in the appropriate space.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
servizio (service)	consegna (delivery)	prezzo (price)	sito (website)	mail	acquisto (purchase)
seriet (earnestness)	spedizione (shipment)	acquistato (purchased)	corriere (carrier)	dopo (after)	assistenza (assistance)
professionalit (professionalism)	ordine (order)	euro	prodotto (product)	email	negozio (shop)
molto (very)	pacco (package)	prodotto (product)	imballo (packaging)	risposta (response)	cura (care)
gentilezza (kindness)	giorni (days)	negozio (shop)	senza (without)	ordine (order)	sito (website)
serieta (earnestness)	dopo (after)	sito (website)	perfetto (perfect)	giorni (days)	prezzi (prices)
professionalita (professionalism)	giorno (day)	samsung	stato (status)	stato (status)	prodotti (products)
ottimo (excellent)	stato (status)	acquisto (purchase)	arrivato (arrival)	ancora (yet)	acquisti (purchases)
consegna (delivery)	ritardo (delay)	trovato (found)	pacco (package)	prodotto (product)	consiglio (advice)
sempre (always)	arrivato (arrival)	spedizione (shipment)	problema (problem)	disponibile (available)	dire (to say)
=..............	=.........	=.........	=........	=..........	=........

English translation is reported in parenthesis.

During the identification, it is possible to change opinion but at the end please fill the most reasonable combination.

Thank you.

Validation Procedure 2: Mapping

In the table below we report and translate each of the five quantitative measures that Trovaprezzi.it uses to ask customers to rate the online merchants after their purchase.

We would like you to tell us which dimension of SERVQUAL each question is measuring. There are no restrictions to your evaluation, simply read the question and tell us which dimension you would assign to that question. If you think the question is not measuring any dimension of SERVQUAL, simply write NONE. If you think the question is potentially measuring more than one dimension please indicate all of them.

Italian	English	SERVUAL Dimension(s)
Facilità di contatto	Ease of contact with merchant	
Facilità di acquisto	Ease of purchasing from merchant	
Facilità di navigazione	Ease of merchant website navigation	
Tempi di consegna	Product delivery speed	
Servizio al cliente	Customer service	

SERVQUAL Dimensions:

•**Reliability**: Ability to perform the promised service dependably and accurately. Examples: 1) Providing services as promised. 2) Performing services right the first time. 3) Providing services at the promised time

•**Responsiveness**: Willingness to help customers and provide prompt and quick service. Examples: 1) Readiness to respond to customers' requests. 2) Convenient business hours, easy to interact with the company. 3) Prompt service to customers. 4) Willingness to help customers.

•**Tangibles**: Appearance of physical facilities, equipment, personnel, and communication materials. Examples: 1) Products are in perfect condition when delivered to the customer. 2) Packaging is visually appealing and in good conditions. 3) Employees who have a neat, professional appearance. 4) Visually appealing materials associated with the service.

•**Assurance**: Knowledge and courtesy of employees and their ability to inspire trust and confidence in customers. Examples: 1) Employees who instill confidence in customers. 2) Making customers feel safe in their transactions. 3) Employees who are consistently courteous. 4) Employees who have the knowledge to answer customer questions.

•**Empathy**: Caring, individualized attention the firm provides its customers. Examples: 1) Giving customers individual attention. 2) Employees who deal with customers in a caring fashion. 3) Having the customer's best interest at heart. 4) Employees who understand the needs of their customers.

Appendix C: Mapping Results

Current system

Evaluation criteria	Customer service expert	Graduate students
Facilità di contatto (Ease of contact with merchant)	Empathy (and maybe assurance)	4 Responsiveness, Empathy, Assurance
Facilità di acquisto (Ease of purchasing from merchant)	None	2 Responsiveness, 3 Assurance
Facilità di navigazione (Ease of merchant website navigation)	Tangibles (just the visual layout of the site)	4 None, Responsiveness, Tangibles
Tempi di consegna (Product delivery speed)	Reliability	4 Reliability, 2 Tangibles, None
Servizio al cliente (Customer service)	Too broad, probably involves all SERVQUAL dimensions, except perhaps tangibles	5 Empathy, 4 Assurance, 3 Responsiveness, Reliability, Tangibles

Table C: Mapping results of company current evaluation system with SERVQUAL dimensions

English translation is reported in parenthesis.

New system

Evaluation criteria	Graduate students
Professionalità e cortesia del personale (Staff professionalism and courtesy)	4 Assurance, 2 Empathy, Responsiveness
Qualità del sito (Website quality)	3 Tangibles, 2 Responsiveness, Empathy
Condizioni del prodotto ricevuto (Received product conditions)	5* Tangibles, Reliability
Affidabilità del merchant (Merchant reliability)	4 Assurance, 3 Reliability, Tangibles, Empathy and Responsiveness
Reperibilità del personale (Staff availability)	5* Responsiveness
Prontezza nel comunicare con il cliente (Readiness to communicate with the customer)	5* Responsiveness, Reliability
Affidabilità dei tempi di consegna (Delivery time trustworthiness)	5* Reliability
Disponibilità verso le richieste del cliente (Availability towards customer' requests)	4 Empathy, 2 Responsiveness

Table D: Mapping results of the new evaluation system proposed

* 5 over 5 respondents, means complete agreement.

References

Archak, N., Ghose, A., and Ipeirotis, P. G. 2011. “ Deriving the Pricing Power of Product Features by Mining Consumer Reviews,” *Management Science* (57: 8), pp. 1485–1509.

Ba, S., Whinston, A. B., and Zhang, H. 2003. “Building trust in online auction markets through an economic incentive mechanism,” *Decision Support Systems* (35:3), pp. 273-286.

Blei, D. M., and Lafferty J. D. 2007. “ A correlated topic model of Science,” *The Annals of Applied Statistics*, (1:1), pp. 17-35.

Blei, D. M. 2012. “Probabilistic topic models,” *Communications of the ACM*, (55 : 4), (April), pp. 77-84.

Blei, D.M., Ng A., and Jordan M .2003. “Latent Dirichlet allocation,” *Journal of machine Learning research*, (3), pp. 993-1022.

Cenfetelli, R. T., Benbasat, I., and Al-Natour, S. 2008. “Addressing the what and how of online services: Positioning supporting-services functionality and service quality for business-to-consumer success,” *Information Systems Research*, (19:2), pp. 161-181.

Delone, W. H., and Mclean, E. R. 2004. “Measuring e-commerce success: Applying the DeLone & McLean information systems success model,” *International Journal of Electronic Commerce*, (9:1), pp. 31-47.

Devaraj, S., Fan, M., and Kohli, R. 2002. “Antecedents of B2C channel satisfaction and preference: validating e-commerce metrics,” *Information Systems Research* (13:3), pp. 316-333.

Duan, W., Cao, Q., Yu, Y., and Levy, S. 2013. “Mining online user-generated content: using sentiment analysis technique to study hotel service quality,” *in System Sciences (HICSS), 2013 46th Hawaii International Conference*, pp. 3119-3128 *on IEEE*.

Feinerer, I., and Hornik, K. 2015. “ tm: Text Mining Package,” R package version 0.6-2. http://CRAN.R-project.org/package=tm

Gefen, D. 2002. “Customer Loyalty in E-Commerce,” *Journal of the Association for Information Systems*, (3:1), pp. 27-51.

Gefen, D., Benbasat, I., and Pavlou, P. 2008. “A research agenda for trust in online environments,” *Journal of Management Information Systems*, (24:4), pp. 275-286.

Ghose, A., and Ipeirotis, P. G. 2006. “Designing ranking systems for consumer reviews: The impact of review subjectivity on product sales and review quality,” *in Proceedings of the 16th Annual Workshop on Information Technology and Systems*, pp. 303-310.

Griffiths, T. and Steyvers, M. 2004. “Finding scientific topics,” *in Proceedings of the National Academy of Science*, (101:suppl 1), pp. 5228-5235.

Gronroos, C. 1982. “Strategic Management and Marketing in the Service Sector,” *Helsingfors: Swedish School of Economics and Business Administration*.

Grönroos, C. 1984. “A service quality model and its marketing implications,” *European Journal of marketing*, (18:4), pp. 36-44.

Gruen, B., and Hornik, K. 2011. “topicmodels: An R Package for Fitting Topic Models,” *Journal of Statistical Software*, (40:13), pp. 1-30. URL http://www.jstatsoft.org/v40/i13/.

Hornik K, Mair P, Rauch J, Geiger W, Buchta C and Feinerer I. 2013. “The textcat Package for n-Gram Based Text Categorization in R,” *Journal of Statistical Software*, (52:6), pp. 1–17. http://doi.org/10.18637/jss.v052.i06.

Hu, N., Pavlou, P. A., and Zhang, J. 2006. “Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of Online word-of-mouth communication,” *in Proceedings of the 7th ACM conference on Electronic commerce,* pp. 324-330. ACM.

Jabr, W., and Zheng, E. 2013. “Know yourself and know your enemy: An analysis of firm recommendations and consumer reviews in a competitive environment,” *MIS Quarterly*. Accepted July.

Jiang, B. J., and Srinivasan, K. 2012. “Pricing and Persuasive Advertising in a Differentiated Market,” unpublished paper, Washington University in St Louis.

Ladhari, R. 2009. “A review of twenty years of SERVQUAL research,” International Journal of Quality and Service Sciences, (1:2), pp. 172-198.

Landis, J. R., and Koch, G. G. 1977. “The measurement of observer agreement for categorical data,” biometrics, pp. 159-174.

Lehtinen, U., and Lehtinen, J. R. 1982. “Service quality: a study of quality dimensions,” Service Management Institute.

Lu B., Ott M., Cardie C. and Tsou B. K. 2011. “Multi-aspect Sentiment Analysis with Topic Models,” *IEEE Computer Society*, pp. 81-88.

Luo, J., Ba, S., and Zhang, H. 2012. “The effectiveness of online shopping characteristics and well-designed websites on satisfaction, ” *MIS Quarterly*, (36:4), pp. 1131-1144.

Madlberger, M., and Nakayama, M. 2013. “On Top Of The World, Down In The Dumps: Text Mining The Emotionality Of Online Consumer,” *Proceedings of European Conference of Information Systems*.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. 2012. “Big data,” The management revolution. *Harvard Business Review*, (90:10), pp. 61-67.

Mudambi, S. M., and Schuff, D. 2010. “What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.Com,” *MIS Quarterly*. (34:1), pp. 185–200.

Murphy, J., Forrest, E. J., Wotring, C. E., and Brymer, R. A. 1996. “Hotel Management and Marketing on the Internet An Analysis of Sites and Features,” *Cornell Hotel and Restaurant Administration Quarterly*, (37:3), pp. 70-82.

Ostrom, A. L., Parasuraman, A., Bowen, D. E., Patricio, L., Voss, C. A. and Lemon, K. 2015. “Service research priorities in a rapidly changing context,” *Journal of Service Research*, (18:2), pp. 127-159.

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*, (49:4), pp. 41-50.

Parasuraman, A., Zeithaml, V.A. and Berry, L.L. 1988. “SERVQUAL: a multi-item scale for measuring customer perceptions of service quality,” *Journal of Retailing*, (64:1), pp. 12-40.

Parasuraman, A., Zeithaml, V.A. and Malhotra, A. 2005. “ES-QUAL a multiple-item scale for assessing electronic service ,” *Journal of Service Research*, (7:3), pp. 213-233.

Petter, S., DeLone, W., and McLean, E. R. 2012. “The past, present, and future of" IS Success," *Journal of the Association for Information Systems*, (13:5), pp. 341-362.

Petter, S., DeLone, W., and McLean, E. R. 2013. “Information Systems Success: the quest for the independent variables,” *Journal of Management Information Systems*, (29:4), pp. 7-62.

Piccoli, G. and Ott, M. 2014. “Sent from my Smartphone: Mobility and time in user-generated content,” *MIS Quarterly Executive*, (13:2).

Piccoli G. and Pigni F. 2013. “Harvesting External Data: The Potential of Digital Data Streams,” *MIS Quarterly Executive* (March), pp. 53-64.

Piccoli G. and Watson R.T. 2008. “Profit from Customer Data by Identifying Strategic Opportunities and Adopting the “Born Digital” Approach,” MIS Quarterly Executive , 7 (3) September, 113-122.

Steyvers, M., and Griffiths, T. 2007. “Probabilistic topic models,” *Handbook of latent semantic analysis*, (427:7), pp. 424-440.

Srivastava, A. N., and Sahami, M. (Eds.). 2009. “Text mining: Classification, clustering, and applications,” CRC Press. pp. 71 -89 (Blei, D. and Lafferty J. D.)

Tan, C. W., Benbasat, I., and Cenfetelli, R. T. 2013. “IT-Mediated Customer Service Content and Delivery in Electronic Governments: An Empirical Investigation of the Antecedents of Service Quality,” *MIS quarterly*, (37:1), pp. 77-109.

Wright, K. B. 2005. “Researching Internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services,” *Journal of Computer-Mediated Communication*, (10:3).

Xu, J. D., Benbasat, I., and Cenfetelli, R. T. 2013. “Integrating service quality with system and information quality: an empirical test in the e-service context, ” *MIS Quarterly*, (37:3), pp. 777-794.