International Journal of Information Management 40 (2018) 132–140

Contents lists available at ScienceDirect

International Journal of Information Management

journal homepage: www.elsevier.com/locate/ijinfomgt

The relative importance of service quality dimensions in E-commerce experiences

B. Palese[a], A. Usai[b,*]

[a] SDEIS Department, E. J. Ourso College of Business, Louisiana State University, United States
[b] Department of economics and business, University of Sassari, Italy

ARTICLE INFO

Keywords:
Online review
Service quality
SERVQUAL
Text mining
Topic model

ABSTRACT

Purpose: The proliferation of socialized data offers an unprecedented opportunity for customer service measurement. This paper addresses the problem of adequately measuring service quality using socialized data.
Design/methodology/approach: The theoretical basis for the study is the widely used SERVQUAL model and we leverage a dataset uniquely suited for the analysis: the full database of online reviews generated on the website of the leading price comparison engine in Italy. Adopting a weakly supervised topic model, we extract the dimensions of service quality from these reviews. We use a linear regression to compare service quality dimensions between positive and negative opinions.
Findings: First, we show that socialized textual data, not just quantitative ratings, provide a wealth of customer service information that can be used to measure service quality. Second, we demonstrate that the distribution of topics in online opinions differs significantly between positive and negative reviews. Specifically, we find that concerns about merchant responsiveness dominate negative reviews.
Practical implications: Our research has important implications for designers of online review systems and marketers seeking novel approaches to the measurement of service quality. Our study shows that evaluation systems designed considering the knowledge extracted directly from customers' review lead to a service quality measurement that not only is theory-based, but also more accurate.
Originality/value: We believe this is the first study to combine the advanced text mining technique of topic modeling and SERVQUAL to extract specific service dimensions from socialized data. Using these advanced techniques, we point to systematic differences between positive and negative customer opinions. We are not aware of any study that has shown these differences with either traditional approaches (i.e., survey data) or modern techniques (e.g. text mining).

1. Introduction

Since its commercialization in 1993, the Internet has dramatically changed people's behavior and decision-making processes. The emergence of the smartphone ecosystem and widespread connectivity has also changed the manner in which we procure goods and services. Individuals' decisions are heavily influenced by other users' personal experiences recorded online in forums and online review websites. At the same time, the variety of products and services available to customers via the online channel continues to increase (Xu, Benbasat, & Cenfetelli, 2013).

Brick and mortar organizations must move online to prevent a loss of market share. However, their lack of technical knowledge and experience with operating online, makes the transition problematic.

Nevertheless, customer service remains a key determinant of e-commerce success (Delone & Mclean, 2004; Wang, 2008) and drives customer satisfaction in online transactions (Cenfetelli, Benbasat, & Al-Natour, 2008; Xu et al., 2013). Information technology enabled customer services challenges traditional views of customers as simple services' receivers, creating opportunities to push the frontier of customer service (Chesbrough & Spohrer, 2006) and to enhance customer satisfaction and loyalty. Companies by harvesting technological innovations can provide high quality and personalized service at reasonable costs (Rust & Miu, 2006).

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

* Corresponding author.
E-mail address: ausai@uniss.it (A. Usai).

https://doi.org/10.1016/j.ijinfomgt.2018.02.001
Received 28 August 2017; Received in revised form 29 January 2018; Accepted 5 February 2018
0268-4012/

quality measurement. Socialized data is data that individuals willingly and knowingly share via digital computer networks (Weigend, 2009). Online reviews are a common form of socialized data, representing spontaneously shared opinions by customers on review platforms (Mudambi & 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 gap is remarkable given the explosion of socialized data. While it has traditionally been difficult to extract useful knowledge from large amounts of information (McAfee, Brynjolfsson, Davenport, Patil, & Barton, 2012) an effective measurement of service quality must be based on customers' experiences (Petter, DeLone, & McLean, 2012).

To contribute to filling this research gap, 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 we are able to extract the five dimensions of SERVQUAL (Parasuraman, Zeithaml, & Berry, 1988) discussed in online reviews. 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, Ba, & Zhang, 2012).

2. Literature review

2.1. Online transactions uncertainty and new sources of information

Quality service is critical in e-commerce to increase channel usage (Devaraj, Fan, & Kohli, 2002), customer loyalty (Gefen, 2002), and customer satisfaction (Cenfetelli et al., 2008; Tan, Benbasat, & Cenfetelli, 2013). Customer service is particularly critical 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, DeLone, & McLean, 2013).

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

Historically, organizations seek to counterbalanced the limitations of the e-commerce environment through website design (Jiang & Srinivasan, 2012), while customers increasingly turn to socialized data to reduce their uncertainty (Piccoli, 2016). 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 & Schuff, 2010).

The computer-mediation of customer service automatically generates data in a digital form (Piccoli & Watson, 2008).

This data can potentially impact not only individual users' decision-making processes but also guide organizations' managers in making strategic decisions (Piccoli & 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 socialized data, the only available option to measure service quality was the use of time consuming customer surveys. 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).

Conversely, customers spontaneously broadcast their opinions about products, services and organizations using opinion platforms and social media. These socialized data offer a wealth of insight to both the firms that are the target of the review as well as other entities, such as competitors, other customers and suppliers.

2.2. Online reviews

Previous research in the business context focused on the effect of online opinions on sales (Chatterjee, 2001; Chen et al. 2004; Chevalier & Mayzlin, 2006; Ghose & Ipeirotis, 2006; Hu, Liu, & Zhang, 2008; Zhu and Zhang, 2010), on trust (Ba & Pavlou, 2002; Pavlou & Gefen, 2004; Pavlou & Dimoka, 2006) and on their helpfulness (Ghose & Ipeirotis, 2011; Mudambi & Schuff, 2010). The literature also focuses on the peer influence of online reviews (Kumar & Benbasat, 2006; Senecal & Nantel, 2004). We are more interested in the impact that they have on purchase decisions and the role that user generated content plays in improving customer service measurement. Kumar and Benbasat (2006) proved that the presence of customer reviews on a website improves customer perception of the usefulness and social presence of the website. Other studies demonstrated their impact on the number of customer visits, their ability to create a community of online shoppers and facilitating consumer decision processes (Dabholkar, 2006; Jiang & Benbasat, 2004; Kohli, Devaraj, & Mahmood, 2004).

It is important to note that 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 thousands of these contributions. Users therefore employ simplifying heuristics, such as examining aggregate quantitative evaluations (i.e., average rating of a product) and the close examination of only a few commentaries (Ghose & Ipeirotis, 2006), when using reviews. Moreover, the distribution of online reviews ratings is bimodal, so the average ratings cannot be considered an accurate measure (Hu, Pavlou, & Zhang, 2006) and an overall neutral rating is not always representative of a neutral opinion (Jabr & Zheng, 2013).

While online opinions have received considerable research attention, there is lack of studies that focused on the role that they play within organizations as part of a customer service measurement system.

The above problems conspire, for both organizations and individual users, to paint an incomplete or misleading picture of customer opinions and experiences. While this is a problem for customers seeking decision-making support in socialized data, it is even more problematic for organizations attempting to measure customers' perception of the service quality using online reviews. We posit that the solution is to leverage the rich text available in socialized data – more specifically by extracting and summarizing the service-specific thematic structure hidden in online reviews. In a previous study, researchers used sentiment analysis to mine the content of online reviews to understand the drivers of users' overall evaluation and content generation (Duan, Cao, Yu, & Levy, 2013). Our work extends this previous attempt, by using the text of online reviews to measure the dimensions of perceived service quality and investigate their effect on customer satisfaction.

3. Theoretical framework

3.1. Service quality

Quality assessment is an important cross-disciplinary area of research in information systems, marketing and operations management.

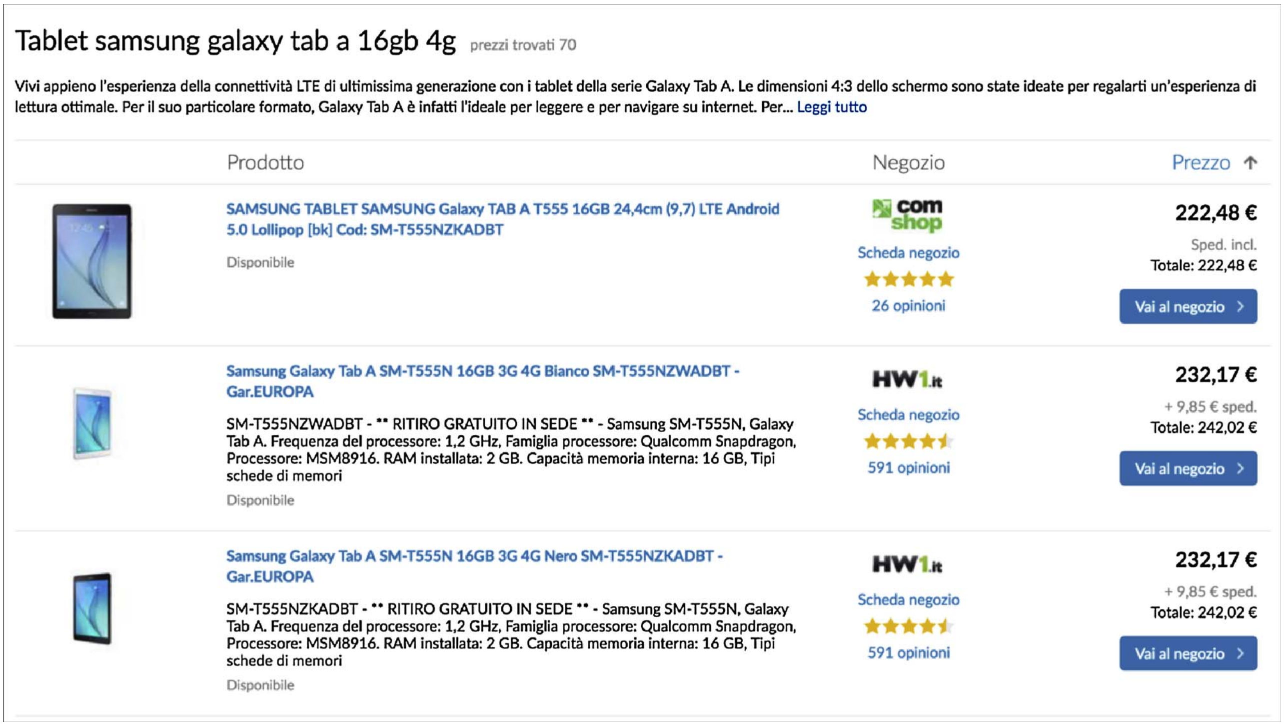

Fig. 1. Search results page.

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 (Grönroos, 1984; Parasuraman, Zeithaml, & Berry, 1985) because they recognize their unique characteristics of intangibility, heterogeneity, and inseparability. The early literature provides varied definitions of service quality. One perspective recognizes technical quality – what the customer is actually receiving from the service – and functional quality – the manner in which the service is delivered (Grönroos, 1982). Another perspective indicates that service is co-produced between a provider and the recipient along three dimensions (Lehtinen & Lehtinen, 1991): physical quality (physical aspects of the service), corporate quality (company's image or profile), and interactive quality (interaction between contact personnel and customers). The SERVQUAL model (Parasuraman et al., 1985) synthesized early work to focus on the difference between initial customer expectation and actual perception. After multiple refinements the SERVQUAL (Parasuraman et al., 1988) coalesced on five dimensions of service quality: 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 (Asubonteng, McCleary, & Swan, 1996; Ladhari, 2009). It 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' perceptions of service quality. In this paper, it is not our intention to enter in the debate on which model developed in literature is better (Lee, Lee, & Yoo, 2000). We note that SERVPERF and SERVQUAL are grounded in the same dimensions.

Rather our focus is on using those same dimensions to investigate their relevance in data that is the text of online reviews socialized by customers. 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 online reviews. We decided to choose the most widely investigated instrument available – namely SERVQUAL – to ground our work (Davies, Baron, Gear, & Read, 1999).

The first objective of our work is to *demonstrate whether the dimensions of the established SERVQUAL model can be extracted directly from the textual component of the online reviews* using topic modeling techniques. Our second objective is to *analyze the relationship between the SERVQUAL dimensions and customer evaluation in online transactions*. 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 online. Previous work has used depth and breadth to measure how much a person cares about an issue (Madlberger & Nakayama, 2013; Piccoli, 2016). Review breadth represents the number of different dimensions discussed in each review by at least one sentence, while review depth is the number sentences used in each review to describe the same dimension (Madlberger & Nakayama, 2013). We adopt this approach as described below.

4. Methodology

4.1. 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 review data (Fig. 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. It is the policy of the price shopping site hosting the reviews that only those customers with verified purchases can write a review assessing their experience with the merchant on the price comparison engine's own website. Thus, our work is immune from the noise associated with fake reviews. The reviews consist of an overall rating of the experience with the merchant 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. It is important to note that customers review the service performance of the merchant, regardless of the product they purchase. As a consequence, our dataset is uniquely suited to answer our research questions. The same could not be said of dataset traditionally used in research based on online reviews (e.g., Amazon) because the focal point of the review is the product, not the provider.

4.2. Data analysis: topic model

With few exceptions (Archak et al. 2011; Duan et al., 2013; Piccoli & 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. More specifically, we are not aware of any research study that has used socialized data or online reviews to extract the dimensions of service quality from the text provided by customers. However, machine learning researchers developed multiple algorithms 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 & Lafferty, 2007). Topic models can extract surprisingly interpretable and useful structures without any "understanding" of language by the computer or any prior training and tagging by humans. 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 & Lafferty, 2007) 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 & 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 & Griffiths, 2007).

4.3. 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 target of the reviews is the service performance of the online merchants listed in the price shopping engine. While they include major vendors (e.g., Amazon) the vast majority of merchants are small regional shops. For these smaller companies with limited brand recognition it is even more important to provide a high quality service and receive good reviews. The database presents the classic J distribution in which positive reviews (58,988) appear one order of magnitude more frequently than the negative reviews (5696). In this section, we consider negative reviews those with one-star rating, while positive reviews are those with five stars.

Online review content is in the form of unstructured textual data, so it is necessary to apply standard preprocessing techniques prior to analysis. We use the R programming language for all analyses (v. 3.3.1).

Through pre-processing, using the *tm* package (Feinerer & Hornik, 2015), we remove singleton words, stop words, numbers, and exclude reviews that were too short – less than 50 words (Lu, Ott, Cardie, & Tsou, 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 & Ott, 2014). We also removed non-Italian reviews using the *textcat* package (Hornik et al., 2013). Upon completion of the pre-processing we were left with 27,117 reviews. The dataset was then tokenized using the *MC_tokenizer* (Feinerer & Hornik, 2015) into unigram and was split into sentences using the *strsplit* function resulting in a total of 122,919 sentences ready for topic modeling.

4.4. 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 & Hornik, 2011). The first phase of multi-aspect sentiment analysis is usually aspect identification. 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 the SERVQUAL dimensions. 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 1).

4.4.1. Topic extraction

To encourage the topic model to learn latent topics that correlate directly with aspects of interest, 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

Table 1
Seed words.

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).

English translation of each seed word is reported in parenthesis.

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".

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, below we report a review from our sample with its English 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).

In this work, we sampled the models for 1000 iterations, with a 500 iterations burn-in and thinning of 10 iterations. We assigned the following value to topic model hyperparameters: $\alpha = 0.01$ and $\eta = 0.1$ (Lu et al., 2011). We tuned the alpha parameter before select the final value. We initially set alpha = 0.1. However, with that alpha the number of undefined sentences was almost 1/4 of the total number of sentences in our corpus. So, we decided to test the algorithm with different alpha values to decrease the number of undefined sentences. The most significant reduction was obtained with alpha = 0.01 (the number dropped from 30855 to 7063). For this reason, we decide to use this alpha in our final model.

4.4.2. Validation

In order to assess the quality of our methodology, we perform a validation of our topic model 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. Five independent raters (graduate students), unaware of the research objectives or the seeding process, classified the topics to provide formal validation of the accuracy or our model. We first provided the context and knowledge necessary to complete the validation. We described in depth the SERVQUAL framework to each rater, including definition and examples for each dimension. Then we provided the raters with the six topics, as described by the ten most frequent terms associated with each one (Table 2). Each rater had to write in the last row the dimension that best resemble each unnamed topic by looking only at the definition of the SERVQUAL dimension and the list of Table 2 terms associated with each of them. Since there were six topics and five SERVQUAL dimensions, raters had to come up with a name for the topic they did not associate with any of the five SERVQUAL dimensions. While they could change their mind as many times as needed during the evaluation, the raters could only label each topic with one dimension.

4.5. Topics' impact evaluation using multiple regression

The second objective of our work is to a*nalyze the relationship between the SERVQUAL dimensions and customer evaluation in online transactions*. To do so, we computed: review breadth and review depth. While review breadth represents the number of different topics (from 0 to 6) discussed in each review by at least one sentence, review depth is the number sentences used in each review to describe the same topic. It provides an indication of each topic's impact on the review. We also computed review length and we used it as a control variable. We then performed a multiple regression analysis to understand how these variables affect the online reviews' overall rating (Eq. (1)). In the next section, we discuss our major findings.

Eq. (1) Multiple regression

$$Rating = \beta_0 + \beta_1\ Review\ length + \beta_2\ Review\ breadth + \beta_3\ Reliability\ depth + \beta_4\ Responsiveness\ depth + \beta_5\ Tangibles\ depth + \beta_6\ Assurance\ depth + \beta_7\ Empathy\ depth + \varepsilon$$

Results

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 & Koch, 1977).

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

IC: 0.858 ± 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 7063 "undefined" sentences (5.75% of the total) that did not unambiguously represent one topic (i.e., no topic had a probability greater than 0.6). Analyzing the remaining sentences, we found that responsiveness (19.13%) and empathy (20.57%) are the preponderant topics in our corpus. On the contrary, tangibles (14.94%) and assurance (13.70%) 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. Moreover, these findings provide an initial indication of what are the topics that customers discuss the most in their reviews.

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. We use multiple regression to achieve this objective. The multicollinearity of our model 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.

The results (Table 3) show that review length has a negative significant effect on overall review rating, while review breadth has a positive significant impact. So, we can infer that the longer the review the lower is the rating associated the review. Moreover, a larger number of topics discussed in one review results to an higher review score.

When we take in consideration the impact of each dimensions, we notice that 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. This indicate that while the presence of some dimensions is beneficial to the overall rate, some others negatively affect it. More specifically, looking at the estimates, the relevance of the responsiveness dimension is clear. In fact, if its depth increases by 1 then the overall rating will decrease by 0.58.

However, to better explain these findings, we examined the topic

Table 2
Topics.

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)
=……………	=………	=………	=………	=………	=………

The words chosen for the seed are marked with *. However, the * was not visible for the raters. English translation is reported in parenthesis.

distributions. Overall, the reviews have a mean breadth of 2.74, indicating that users discuss, on average, between two and three different topics per review. Interestingly, responsiveness is discussed in the lowest percentage among topics (Table 4). However, there are stark differences in depth by topic. Reviews discussing responsiveness are split about evenly between those with depth of 1 (52.20%) and those addressing responsiveness with more than one sentence (47.80%). A quarter of reviews addressing responsiveness have depth greater than two (25.74%). Conversely, the other dimensions only have around 10% of reviews with more than two sentences dedicated to the same service quality dimension (reliability: 7.96%, tangibles: 12.01%, assurance: 6.44%, empathy: 14.25%). 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.

It is the focus of negative reviews on responsiveness that explains the difference in distribution by topic (Fig. 2). While negative reviews are only one fourth of the sample, they are dominated by sentences focusing their discussion of poor service quality on the responsiveness dimension. Responsiveness is discussed in only a quarter of positive reviews (24.11%) while almost all negative opinions address it (89.03%). Further, responsiveness is the only topic that presents a U shape (instead of the typical J distribution of review valence in our dataset). 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.

6. Managerial implication

The findings of our study 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 the SERVQUAL dimensions have different distributions in terms of rating and depth. Companies looking to improve their service quality need to consider these differences. For example, when customers discuss the responsiveness of the merchants, they emphasize this aspect of the service experience disproportionately more than any other topic.

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, we asked to map SERVQUAL dimensions to the evaluation criteria composing the system currently 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 5, Appendix A) 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, since it is highly discussed in negative reviews and it has a negative impact on the overall rating. 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 6, Appendix A) 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. The measurement accuracy can improve if companies design these systems by including in the evaluation elements that are relevant for their customers.

7. Limitations and conclusions

As with any study of this nature we acknowledge some limitations. While our findings provide useful insights, they are affected by the selection of the service quality model. 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

Table 3
Multiple regression results.

Coefficients	Estimate	Std. Error	t value	Pr(> |t|)	VIF
(Intercept)	4.5019900	0.0202114	222.745	< 2e-16***	
Review length	−0.0040038	0.0001786	−22.418	< 2e-16***	1.658795
Review breadth	0.0337213	0.0100528	3.354	0.000796 ***	2.496818
Reliability depth	0.2196883	0.0094984	23.129	< 2e-16***	1.411930
Responsiveness depth	−0.5816510	0.0064059	−90.800	< 2e-16***	1.670625
Tangibles depth	−0.1392828	0.0085903	−16.214	2.6e-12***	1.473743
Assurance depth	0.3135167	0.0102206	30.675	< 2e-16***	1.395727
Empathy depth	0.1325495	0.0077498	17.104	< 2e-16***	1.428999
(mean VIF)					1.648091

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

Table 4
Number of review divided by number of sentences associated to each topic.

	Depth by SERVQUAL dimension							Reviews	Dimension Proportion
	0	1	2	3	4	5	> 5		
reliability	13801	9372	2884	810	178	57	15	13,316	49%
responsiveness	16359	5616	2373	1245	681	398	445	10,758	40%
tangibles	16036	7208	2542	808	314	126	83	11,081	41%
assurance	15588	8245	2542	601	114	22	5	11,529	43%
empathy	12485	8814	3728	1388	459	159	84	14,632	54%

accurately capturing the topics affecting customers' evaluation. Moreover, we only extracted topics representing the SERVQUAL dimensions. So, another extension would be to control other aspects that can impact the overall customer experience (price, product quality etc.). Furthermore, our dataset includes only review in Italian related to Italian merchants. We plan to use the same methodology on datasets that contain English reviews about international merchants to improve the generalizability of the study.

Despite the above limitations, 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 & Zheng, 2013). 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 large 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, this methodology enables the systematized assessment of service quality systems able to reliably measure all the aspects that influence customer evaluations directly from customers' opinion. Companies by harvesting online reviews content can gather relevant information to assess the quality of the service and their customers satisfactions. 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 & Ipeirotis, 2006) can improve customer experience online.

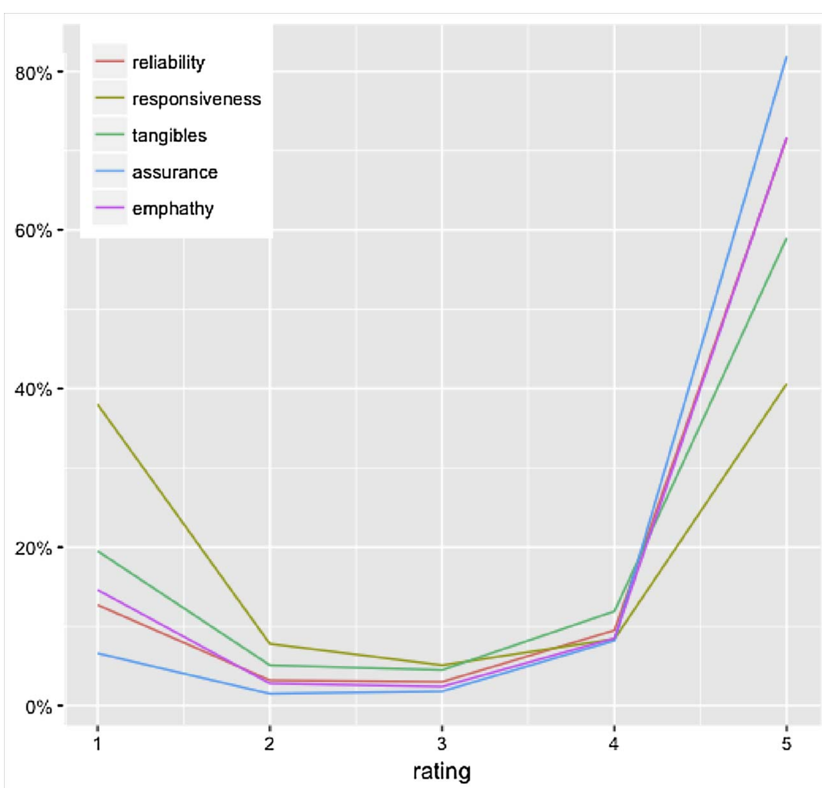

Fig. 2. Topic distribution among reviews' rating.

Appendix A. Mapping Results

Current system.

See Table 5.

Table 5
Mapping results of company current evaluation system with SERVQUAL dimensions.

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

English translation is reported in parenthesis.

New system.

See Table 6.

Table 6
Mapping results of the new evaluation system proposed.

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

*5 over 5 respondents, means complete agreement.

References

Archak, N., Ghose, A., & Ipeirotis, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management Science, 57*(8), 1485–1509.

Asubonteng, P., McCleary, K. J., & Swan, J. E. (1996). SERVQUAL revisited: a critical review of service quality. *Journal of Services Marketing, 10*(6), 62–81.

Ba, S., & Pavlou, P. (2002). Evidence of the Effect of Trust Building Technology in Electronic Markets: Price Premiums and Buyer Behavior. *MIS Quarterly, 26*(3), 243–268.

Ba, S., Whinston, A. B., & Zhang, H. (2003). Building trust in online auction markets through an economic incentive mechanism. *Decision Support Systems, 35*(3), 273–286.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics, 1*(1), 17–35.

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

Cenfetelli, R. T., Benbasat, I., & 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), 161–181.

Chatterjee, P. (2001). Online review: do consumers use them. *Advances in Consumer Research, 28*, 129–133.

Chen, P.-Y., Wu, S.-Y., & Yoon, J. (2004). The impact of online recommendations and consumer feedback on sales. *Proceedings of the International Conference on Information Systems,* 711–724.

Chesbrough, H., & Spohrer, J. (2006). A research manifesto for services science. *Communications of the ACM, 49*(7), 35–40.

Chevalier, J., & Mayzlin, D. (2006). The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research, 43*(3), 345–354.

Dabholkar, P. (2006). Factors influencing consumer choice of a'Rating web site': An experimental investigation of an online interactive decision aid. *Journal of Marketing Theory and Practice, 14*(4), 259–273.

Davies, B., Baron, S., Gear, T., & Read, M. (1999). Measuring and managing service quality. *Marketing Intelligence & Planning, 17*(1), 33–40.

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

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

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

Feinerer, I., & Hornik, K. (2015). tm: Text mining package. *R Package Version 0.6-2* [Available at http://CRAN.R-project.org/package=tm (accessed 12 May 2017)].

Gefen, D., Benbasat, I., & Pavlou, P. (2008). A research agenda for trust in online environments. *Journal of Management Information Systems, 24*(4), 275–286.

Gefen, D. (2002). Customer loyalty in E-Commerce. *Journal of the Association for Information Systems, 3*(1), 27–51.

Ghose, A., & Ipeirotis, P. G. (2006). Designing ranking systems for consumer reviews: the impact of review subjectivity on product sales and review quality. *Proceedings of the 16th Annual Workshop on Information TechNo.logy and Systems* (pp. 303–310).

Ghose, A., & Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics Knowledge and Data Engineering. *IEEE Transactions on, 23*(10), 1498–1512.

Grönroos, C. (1982). *Strategic management and marketing in the service sector, working paper.* Helsinki, Finland: Helsingfors: Swedish School of EcoNo.mics and Business Administration.

Grönroos, C. (1984). A service quality model and its marketing implications. *European Journal of Marketing, 18*(4), 36–44.

Griffiths, T., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Science, 101*(1), 5228–5235.

Gruen, B., & Hornik, K. (2011). Topicmodels: an R package for fitting topic models. *Journal of Statistical Software, 40*(13), 1–30 [Available at http://www.jstatsoft.org/v40/i13/ accessed 12 May 2017].

Hornik, K., Mair, P., Rauch, J., Geiger, W., Buchta, C., & Feinerer, I. (2011). The textcat package for n-gram based text categorization in r. *Journal of Statistical Software, 52*(6), 1–17. http://dx.doi.org/10.18637/jss.v052.i06 [accessed 12 May 2017].

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

Hu, N., Liu, L., & Zhang, J. J. (2008). Do online reviews affect product sales? The role of reviewer characteristics and temporal effects. *Information Technology and Management, 9*(3), 201–214.

Jabr, W., & Zheng, E. (2013). Know yourself and kNo. w your enemy: An analysis of firm recommendations and consumer reviews in a competitive environment. *MIS Quarterly* [Accepted July].

Jiang, Z., & Benbasat, I. (2004). Virtual product experience: Effects of visual and functional control of products on perceived diagnosticity and flow in electronic shopping. *Journal of Management Information Systems, 21*(3), 111–147.

Jiang, B. J., & Srinivasan, K. (2012). *Pricing and persuasive advertising in a differentiated market.* Washington University in St Louis [unpublished paper].

Kohli, R., Devaraj, S., & Mahmood, M. A. (2004). Understanding determinants of online consumer satisfaction: A decision process perspective. *Journal of Management Information Systems, 21*(1), 115–135.

Kumar, N., & Benbasat, I. (2006). The influence of recommen- dations on consumer reviews on evaluations of websites. *Information Systems Research, 17*(4), 425–439.

Ladhari, R. (2009). A review of twenty years of SERVQUAL research. *International Journal of Quality and Service Sciences, 1*(2), 172–198.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics,* 159–174.

Lee, H., Lee, Y., & Yoo, D. (2000). The determinants of perceived service quality and its relationship with satisfaction. *Journal of Services Marketing, 14*(3), 217–231.

Lehtinen, U., & Lehtinen, J. R. (1991). Two approaches to service quality dimensions. *Service Industries Journal, 11*(3), 287–303.

Lu, B., Ott, M., Cardie, C., & Tsou, B. K. (2011). Multi-aspect sentiment analysis with topic models. *IEEE Computer Society,* 81–88.

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

Madlberger, M., & 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), 61–67.

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon.Com. *MIS Quarterly, 34*(1), 185–200.

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

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1985). A conceptual model of service quality and its implications for future research. *Journal of Marketing, 49*(4), 41–50.

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

Pavlou, P. A., & 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, 17*(4), 392–414.

Pavlou, P., & Gefen, D. (2004). Building effective online marketplaces with institution-based trust. *Information Systems Research, 15*(1), 37–59.

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

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

Piccoli, G., & Ott, M. (2014). Sent from my Smartphone: Mobility and time in user-generated content. *MIS Quarterly Executive, 13*(2), 147–157.

Piccoli, G., & Pigni, F. (2013). Harvesting external data: The potential of digital data streams. *MIS Quarterly Executive, March*, 53–64.

Piccoli, G., & 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.

Piccoli, G. (2016). Triggered essential reviewing: the effect of techNo. logy affordances on service experience evaluations. *European Journal of Information Systems, 25*(6), 477–492.

Rust, R. T., & Miu, C. (2006). What academic research tells us about service. *Communications of the ACM, 49*(7), 49–54.

Senecal, S., & Nantel, J. (2004). The influence of online product recommendations on consumers' online choices. *Journal of Retailing, 80*(2), 159–169.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of Latent Semantic Analysis, 427*(7), 424–440.

Tan, C. W., Benbasat, I., & 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), 77–109.

Wang, Y. S. (2008). Assessing e-commerce systems success: A respecification and validation of the DeLone and McLean model of IS success. *Information Systems Journal, 18*(5), 529–557.

Weigend, A. (2009). *The social data revolution s.blogs.harvardbusiness.org, 20 05.* [Avalable at https://hbr.org/2009/05/the-social-data-revolution (accessed 12 May 2017)].

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., & Cenfetelli, R. T. (2013). Integrating service quality with system and information quality: an empirical test in the e-service context. *MIS Quarterly, 37*(3), 777–794.

Zhu, F., & Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing, 74*(2), 133–148.