Online Reviews as a Measure of Service Quality

Full Paper

Introduction

Since its commercialization in 1993, the Internet has dramatically changed people's behavior. Today we communicate by instant messaging, we share pictures on social networks, we tag our geolocation. Perhaps more fundamentally the Internet has altered the manner in which people make decisions – from product purchasing (Wolfinbarger and Gilly, 2001) to hotel selection (Murphy et al., 1996) to finding love (Ellison et al., 2006). Individuals' decisions are today heavily influenced by other users' personal experiences recorded online in forums and online review websites.

Previous research in the business context focused on the effect of online opinions on sales (Ghose and Ipeirotis, 2006; Hu et al., 2008), on trust (Ba and Pavlou, 2002; Pavlou and Gefen, 2004; Pavlou and Dimoka, 2006) and on their helpfulness (Mudambi and Schuff, 2010). The literature also focuses on the peer influence of online reviews (Kumar and Benbasat, 2006). 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 (Jiang and Benbasat 2007).

While online opinions have received considerable research attention, to our knowledge, no study has focused on the role that they play within organizations as part of a customer service measurement system. A recent exception is offered by Duan et al. (2013). Our work extends this previous attempt to measure the dimensions of customer service using the text of online reviews. Specifically, we use weakly supervised topic modeling (Lu et al. 2011), to measure the dimensions of perceived service quality and to measure its effect on customer satisfaction.

Theoretical Framework

Individuals' opinions are a crucial source of insight for companies seeking to evaluate service quality and measure customer satisfaction (Baker and Crompton, 2000). These opinions are increasingly "socialized" by users, making them available to organizations. Socialized data is data that individuals knowingly and willingly share. Online reviews are a common form of socialized data. Specifically they are feedback spontaneously shared by customers on appropriate review platforms (Mudambi and Schuff, 2010). Socialized data differ substantially from traditional information sources for customer service assessment, such as customer surveys or word of mouth. Socialized data, by definition, is 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). In online reviews the distribution of ratings is bimodal, so the average ratings cannot be considered an accurate measure (Hu et al., 2006). Also reading just a few recent reviews is unlikely to yield accurate perception of a service quality.

Service quality

Quality assessment is an important cross-disciplinary area of research. 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.

Service quality comprises technical quality – what the customer is actually receiving from the service – and functional quality – the manner in which the service is delivered (Gronroos, 1982). 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).

Service quality measurement

Parasuraman et al. (1985), conducted focus groups with customers and in-depth interviews with executives to discover a set of discrepancies, called gaps, regarding executive perceptions of service quality and the tasks associated with service delivery to consumers. In particular, the customer gap is the most critical. It measures the distance between customers' initial expectation and final service's perception. The criteria used to evaluate the quality of the service don't vary across industries (Parasuraman et al., 1985) and fall into five specific dimensions: namely, reliability, responsiveness, tangibles, assurance and empathy (Parasuraman et al., 1988).

New Approaches to the Measurement of Service Quality

Until the emergence of socialized data it was quite difficult to measure service quality. Surveys were the main mean of collection of data necessary to assess the quality of a service. However, customers are increasingly overwhelmed by company communications (e.g., email, phone calls, robo-calls) eliciting 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).

IT advances have profoundly changed the way information is transmitted, and have transcended the traditional limitations of word-of-mouth. Consumers can now easily and freely access information and exchange opinions on companies, products and services on an unprecedented scale in real time. The rise of Web 2.0 first, and the shift to the mobile platform later, enabled the birth of a vast number 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 (usually ranging from one to five stars) and open-ended customer-authored comments about the product (Mudambi and Shuff, 2010). The computer-mediation of customer service automatically generates data in a digital form (Piccoli and Watson, 2008). Digital data streams can potentially impact not only single users' decision-making processes but also guide organizations' managers in making strategic decisions (Piccoli and Pigni, 2013). For example, they can be used to better understand market reactions to companies' current offers and subsequently feed this information into their product development and quality control processes (Dellarocas, 2003). However, it is necessary that these data are appropriately streamed. Online reviews are an example of a digital data stream that can be effectively collected by companies and/or researchers allowing a continuous monitoring of customers' satisfaction. By some accounts customers post 200 new contributions every minutes (Fact sheet, TripAdvisor) and generate more than 30 million reviews last year (Fact sheet, Yelp).

The first objective of our exploratory work is to *demonstrate whether the dimensions of the SERVQUAL model can be extracted from socialized data*. The second objective of our work is to analyze *which are the SERVQUAL dimensions that mostly influence the overall customer evaluation?*

Methodology

Data analysis: Topic model

The first objective of our research is to *demonstrate whether the dimensions of the SERVQUAL model can be extracted from socialized data*. With few exceptions (Arckack 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. Early approaches to automatically extract and interpret review text have focused on determining either the overall polarity or the sentiment rating of a review. But considering coarse overall ratings fails to represent the multiple dimensions of service quality. Topic modeling, an innovative 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. With these estimates it is possible to perform information

retrieval or text mining tasks on the corpus. In our analysis we use a weakly supervised approach to topic modeling using Gibbs-sampling. A Markov chain is constructed, 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 reviews refer to different online shops available for price comparison on the company website. The database presents a J distribution in which positive reviews (58,988) are more than ten times the negatives ones (5,696). In this section, we consider negative reviews those with one-star rating while positive reviews are those with five stars.

Through standard pre-processing we remove singleton words, stop words 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 the dataset was tokenized into unigram and it was split into sentences resulting in a total of 27,117 reviews and 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 perform multi-aspect sentence labeling using the *topicmodels* packages in R (Gruen and Hornik, 2011). This approach utilizes only minimal prior knowledge, in the form of seed words, to enforce a direct correspondence between topics and aspects. The first phase of multi-aspect sentiment analysis is usually aspect identification, in this paper we used the dimensions of SERVQUAL as aspects and we selected words (using only nouns) that are able to capture their essence directly from the vocabulary of our corpus.

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 label each sentence with the most prevalent topic. 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 seed words for the LDA model represent 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 $\beta \sim Dir(\{\eta + C\_w\}: w \in Seed)$, 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).
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 created the seeding with frequent corpus terms that adequately capture the meaning of SERVQUAL dimensions. The pseudo count C_w for seed words was heuristically set to be 3000 (~ 10% of the number of reviews as in 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 role of the seeding is to improve the accuracy of sentence labels in our corpus. The six labels associated with each sentence are: reliability, responsiveness, tangibles, assurance, empathy and "background".

The second objective of our work is to *identify the SERVQUAL dimensions that mostly influence the overall customer evaluation*. To do so we computed two new variables: review breadth and review depth (Piccoli and Ott, 2014). Breadth represents the number of *different topics* (0 to 6) discussed in each review by at least one sentence. Review depth is the number sentences (1 to ∞) used in each review to describe the same topic. We then performed a multiple regression analysis to understand how these variables affect the online reviews' overall rating.

Results

The output of topic modeling is a set of k topics. 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, so it needed to be confirmed through independent validation. The participants to the validation did not have information regarding the seeding and the model in general and they were unaware of the research objectives. To validate the model, respondents had to correctly assign SERVQUAL dimensions to unnamed topics represented by the ten most frequent words. 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 show that agreement is deemed almost perfect (Landis and Koch, 1977).

We first conducted an analysis at the sentence level. We removed 30,742 sentences that did not unambiguously represent one topic (i.e., no topic had a probability greater than 0.6). Responsiveness (20%) and empathy (22%) are the preponderant topics in our corpus. On the contrary, tangibles (12%) and assurance (13%) are discussed less often. 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 provided by the merchants. This is a review-level analysis we perform by estimating the model below:

$$Overall\ rating = \beta_0 + \beta_1 Sentence\ number + \beta_2\ Review\ lenght + \beta_3\ Review\ breadth + \beta_4\ Reliability\ depth + \beta_5\ Responsiveness\ depth + \beta_6\ Tangibles\ depth + \beta_7\ Assurance\ depth + \beta_8\ Empathy\ depth + \varepsilon$$

The results (Table 2) show that the number of sentences and review breadth have a positive and significant impact on the overall review ratings, while review length negatively affects it. Among topics' depth, only the depth of responsiveness has a significant negative impact on the rating, with reliability and tangibles being marginally significant. These three dimensions are negatively related to overall rating – thus indicating that negative reviews have greater depth.

Coefficients	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
Intercept	5.2269862	0.0220867	237.658	< 2e-16 ***
Sentences number	0.0376109	0.0059447	6.327	2.54e-10 ***
Review length	-0.0130836	0.0002216	-59.048	< 2e-16 ***
Review breadth	0.0422937	0.0163086	2.593	0.00951 **
Reliability depth	-0.0283484	0.0156067	-1.816	0.06932 .
Responsiveness depth	-0.0317428	0.0126027	-2.519	0.01178 *
Tangibles depth	-0.0311170	0.0183348	-1.697	0.08968 .
Assurance depth	-0.0289660	0.0182738	-1.585	0.11295
Empathy depth	-0.0175162	0.0159006	-1.102	0.27064

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

Table 2: Multiple regression results

To better explain these findings we examine the topic distributions. Overall the reviews have 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 1). 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 or reviews addressing responsiveness have depth greater than two (24.32%). Conversely, the other dimensions have only 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. People are not "superficial" about responsiveness.

	Depth by SERVQUAL dimension							Reviews	Dimension Proportion
	0	1	2	3	4	5	> 5		
reliability	15441	8496	2422	566	157	27	8	11676	43%
responsiveness	17811	5003	2040	1097	542	325	299	9306	34%
tangibles	19125	5913	1443	422	157	36	21	7992	29%
assurance	17738	7121	1870	330	48	9	1	9379	35%
empathy	14031	8536	3004	1034	353	106	53	13086	48%
background	16054	8602	2020	343	78	13	7	11063	41%

Table 1: 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 1). 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 curve (instead of the typical J distribution). Moreover, in absolute term the negative peak is even higher than the positive one.

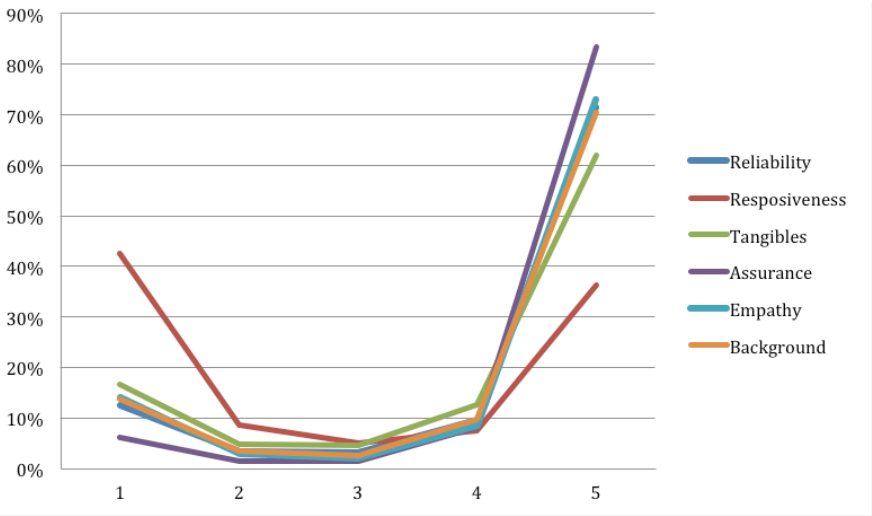

Figure 1: Topic distribution among reviews' rating

Practical Implications

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 current evaluation system. The results of this validation showed that some of the system evaluation criteria are too broad while others are unable to capture any of the topics. Moreover, none of them is able

to accurately measure responsiveness, the most influential topic in our findings. We proposed a new evaluation system to the price shopping company.

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. Moreover, service quality evaluation systems should be able to map with reviews' topic content in order to improve customers experience and to increase measurement accuracy. Companies that want to achieve high customers' satisfaction 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, should be used to extract meaning from the huge amount of socialized data that can be harvested. 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 customers that generally take a decision based on the quantitative rating of inaccurate criteria, and to organization gaining real time knowledge of customers' opinions.

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 to broaden the study to different industries.

References

Archak, N., Ghose, A., and Ipeirotis, P. G. 2011. " Deriving the Pricing Power of Product Features by Mining Consumer Reviews," Management Science Vol. 57, No. 8, pp. 1485–1509.

Ba, S., and 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.

Baker, D. A., and Crompton, J. L. 2000. "Quality, satisfaction and behavioral intentions," Annals of tourism research, 27(3), 785-804.

Blei D. M. and 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

Blei D.M., Ng A., and Jordan M .2003. "Latent Dirichlet allocation," Journal of machine Learning research, 3, 993-1022.

Dellarocas, C. 2003. ”The digitization of word of mouth: promise and challenges of online feedback mechanisms, ” Management Science 49 (10) 1407–1424.

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 on (pp. 3119-3128). IEEE.

Ellison, N., Heino, R., and Gibbs, J. 2006. “Managing impressions online: Self-presentation processes in the online dating environment,” Journal of Computer-Mediated Communication, 11(2), 415-441.

Fact Sheet - TripAdvisor. Retrieved March 01, 2016, from http://www.tripadvisor.com/PressCenter-c4-Fact_Sheet.html

Fact Sheet -Yelp. Retrieved March 01, 2016, from http://www.yelp.com/factsheet

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).

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), 36-44.

Jiang, Z., and 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.

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

Kumar, N., and Benbasat, I. 2006. “The Influence of Recommen- dations on Consumer Reviews on Evaluations of Websites,” Information Systems Research, 17(4), 425-439.

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.

Hu, N., Liu, L., and 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.

Landis, J. R., and Koch, G. G. 1977. “The measurement of observer agreement for categorical data,” biometrics, 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, 81-88.

Mudambi, S. M., and Schuff, D. 2010. “What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.Com,” MIS Q. (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), 70-82.

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), 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), 12-40.

Pavlou, P. A. , and Dimoka, A. 2006. "The Nature and Role of Feedback Text Comments in Online Marketplaces: Implications for Trust Building, Price Premiums, and Seller Differentiation," Information Systems Research (17:4), pp. 392–414.

Pavlou, P., and Gefen, D. 2004. "Building Effective Online Marketplaces with Institution-Based Trust," Information Systems Research, 15(1), 37-59.

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

Piccoli G. and Pigni F. 2013. "Harvesting External Data: The Potential of Digital Data Streams," MIS Quarterly Executive (March), 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), 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.)

Wolfinbarger, M., and Gilly, M. C. 2001. "Shopping online for freedom, control, and fun," California Management Review, 43(2), 34-55.

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), 00-00.