Designing Novel Review Ranking Systems: Predicting Usefulness and Impact of Reviews

Anindya Ghose
aghose@stern.nyu.edu

Panagiotis G. Ipeirotis
panos@nyu.edu

Department of Information, Operations, and Management Sciences
Leonard N. Stern School of Business, New York University

ABSTRACT

With the rapid growth of the Internet, users' ability to publish content has created active electronic communities that provide a wealth of product information. Consumers naturally gravitate to reading reviews in order to decide whether to buy a product. However, the high volume of reviews that are typically published for a single product makes it harder for individuals to locate the best reviews and understand the true underlying quality of a product based on the reviews. Similarly, the manufacturer of a product needs to identify the reviews that influence the customer base, and examine the content of these reviews. In this paper we propose two ranking mechanisms for ranking product reviews: a consumer-oriented ranking mechanism ranks the reviews according to their expected helpfulness, and a manufacturer-oriented ranking mechanism ranks the reviews according to their expected effect on sales. Our ranking mechanism combines econometric analysis with text mining techniques and with subjectivity analysis in particular. We show that subjectivity analysis can give useful clues about the helpfulness of a review and about its impact on sales. Our results can have several implications for the market design of online opinion forums.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*text analysis*; H.2.4 [**Database Management**]: Systems—*Textual databases*; H.2.8 [**Database Applications**]: Data mining; J.4 [**Social And Behavioral Sciences**]: Economics

General Terms

Algorithms, Measurement, Economics, Experimentation

Keywords

consumer reviews, econometrics, electronic commerce, electronic markets, opinion mining, product review, sentiment analysis, text mining, user-generated content, Web 2.0

ICEC'07, August 19–22, 2007, Minneapolis, Minnesota, USA.

1. INTRODUCTION

In offline markets, consumers' purchase decisions are heavily influenced by word-of-mouth. With the rapid growth of the Internet these conversations have migrated in online markets, creating active electronic communities that provide a wealth of product information. Consumers now rely on online product reviews, posted online by other consumers, for their purchase decisions [3]. Reviewers contribute time, energy, and other resources, enabling a social structure that provides benefits both for the users and the companies that host electronic markets. Indeed, the provision of a forum facilitating social exchanges in the form of consumer product reviews is an important part of many electronic markets, such as Amazon.com.

Unfortunately, a large number of reviews for a single product may also make it harder for individuals to evaluate the true underlying quality of a product. This is especially true when consumers consider the average rating of a product to make decisions about purchases or recommendations. Recent work has shown that the distribution of an overwhelming majority of reviews posted in online markets is bimodal [13]. Reviews are either allotted an extremely high rating or an extremely low rating. In such situations, the average numerical star rating assigned to a product may not convey a lot of information to a prospective buyer. Instead, the reader has to read the actual reviews to examine which of the positive and which of the negative aspect of the product are of interest. In these cases, buyers may naturally gravitate to reading a few reviews in order to form a decision regarding the product. Similarly, manufacturers want to read the reviews to identify what elements of a product affect sales most.

In this paper, we propose two ranking mechanisms for ranking product reviews: a consumer-oriented ranking mechanism ranks the reviews according to their expected helpfulness, and a manufacturer-oriented ranking mechanism ranks the reviews according to their expected effect on sales. So far, the best effort for ranking reviews for consumers comes in the form of peer reviewing in the review forums, where customers give helpful votes to other reviews. In digital markets, individuals use peer ratings to confirm that other reviewers are member in good standing within the community [8]. Unfortunately, the helpful votes are not a useful feature for ranking recent reviews: the helpful votes are accumulated over a long period of time, and hence cannot be used for review placement in a short- or medium-term time

frame.

As a major contribution, our techniques examine the actual text of the review to identify which review is expected to have the most impact. We show that the actual style of the review plays an important role in determining the impact of the review: reviews that confirm the information contained in the product description are the more important for feature-based[1] products, while reviews that give a more subjective point of view are more important for experience goods, such as movie DVDs. Similarly, we show that the style of a review can also influence sales. Contrary to the intuition, we observed that reviews that are considered helpful by the users are not necessarily influential, and vice versa.

Based on such results, we posit that the actual textual content of each review plays an important role in influencing consumer purchase decisions and thereby affecting actual sales of the product. We investigate the veracity of this theory and quantify the extent to which textual content of each review affects product sales on a market such as Amazon. While prior work in computer science has extensively analyzed and classified sentiments in online opinions [12, 14, 15, 17, 21, 23], they have not examined the economic impact of the reviews.

The rest of the paper is structured as follows. First, in Section 2, we describe our data set. Then, in Section 3, we give the details of our algorithmic approach for analyzing the subjectivity of a review. In Section 4, we present our econometric analysis that uses the results of our text mining algorithm. Section 5 has the details of a content analysis we perform using independent coders to validate our empirical results. Finally, Section 6 discusses related work and Section 7 provides some additional discussion and concludes the paper.

2. DATA

To conduct our study, we created a panel data set of products from Amazon.com, using publicly available information about product prices and sales rankings. We gathered the information using automated Java scripts that access and parse HTML and XML pages, over the period of March 2005–May 2006. In our data set, we had a set of different products belonging to different categories. Specifically, we have the following categories: DVDs, audio and video players, videogames, computers, PDAs software, and digital cameras. However, for brevity we present our empirical analysis using two product categories: (i) audio and video players, and (ii) digital cameras. For each of the products in our data set, we collected two sets of information.

Product and Sales Data: The first part of our data set consists of product specific characteristics, collected over time. We include the list price of the product, its Amazon retail price, its Amazon sales rank (which serves as a proxy for units of demand, as described further later), and the date the product was released into the market. We also have some secondary market data such as the number of used versions of that good that are available for sale and the minimum price of the used good.

Reviews: The second part of out data set consists of the details of product reviews. We collected all reviews of a product chronologically since the product was released into the market until the end of the time period of our data collection. Amazon has a voting system whereby community members can provide helpful votes to rate the reviews of other community members. For each review, we retrieve the actual textual content of the review, the rating of a product given by the reviewer, the total number of "helpful votes" received by the review, and the total number of votes that were posted for that review. The rating that a reviewer allocates to a review is denoted by a number of stars on a scale of 1-5.

The summary statistics of the data are given in Table 1.

3. ESTIMATING THE SUBJECTIVITY OF A REVIEW

Our approach is based on the hypothesis that the actual text of the review matters. Previous text mining approaches focused on extracting automatically the polarity of the review [4, 6, 11, 12, 14, 18–24]. In our setting, the numerical rating score already gives the (approximate) polarity of the review,[2] so we look in the text to extract features that are not possible to observe using simple numeric ratings. In particular, we are interested to examine what types of reviews affect most sales and what types of reviews are most helpful to the users. We assume that there are two types of reviews, from the stylistic point of view. There are reviews that list "objective" information, listing the characteristics of the product, and giving an alternate product description that confirms (or rejects) the description given by the merchant. The other types of reviews are the reviews with "subjective," sentimental information, in which the reviewers give a very personal description of the product, and give information that typically does not appear in the official description of the product.

As a first step towards understanding the impact of the textual content of the reviews on product sales, we rely on existing literature of subjectivity estimation from computational linguistics [19]. Specifically, Pang and Lee described a technique that identifies which sentences in a text convey objective information, and which of them contain subjective elements. Pang and Lee applied their techniques in a movie review data set, in which they considered as objective information the movie plot, and as subjective the information that appeared in the reviews. *In our scenario, objective information is considered the information that also appears in the product description, and subjective is everything else.*

This resulted in a training set with two classes of documents:

- A set of "objective" documents that contains the product descriptions of each of the 1,000 products in our data set.
- A set of "subjective" documents that contains randomly retrieved reviews.

[1]Feature-based products can be viewed as consisting of various bundles of a small number of characteristics or basic attributes [2]. For instance, a digital camera can be decomposed to characteristics such as weight (w), megapixels (c), and storage capacity (p). Digital cameras, appliances, music players, and so on, fall into this category in contrast to experience goods such as movies and books that do not have clear utilitarian characteristics.

[2]We should note, though, that the numeric rating does not capture all the polarity information that appears in the review [1].

Variable	Obs.	Mean	Std. Dev.	Min	Max
AvgProb	18720	.58396	.04495	.37	.8297
DevProb	18720	.04756	.02378	0	.1807
Sales Rank	18628	7667.42	51039.42	0	2090308
Rating	18720	3.8563	1.4141	1	5
Helpful Votes	18720	6.3432	13.873	0	706
Total Votes	18720	9.6248	16.359	0	847
Reviews	18616	138.421	202.24	0	1339
Amazon Price	15108	76.6312	162.73	0	7999.99
Used Price	16318	116.033	181.58	0	7999
Num. of Used Goods	12057	39.8082	38.91	0	241
Sentences	18720	10.3533	10.42	1	160
Log(Length)	18720	4.5512	0.527	2.681	8.1373
Log(Elapsed Date)	17000	5.1225	1.095	0	7.6338
Moderate	18721	.09337	.2909	0	1

Table 1: Descriptive statistics based on all product categories.

Since we deal with a rather diverse data set, we constructed separate subjectivity classifiers for each of our product categories. We trained the classifier using a Dynamic Language Model classifier with n-grams ($n = 8$) from the LingPipe toolkit[3].

After constructing the classifiers for each product category, we used the resulting classification models in the remaining, *unseen* reviews. Instead of classifying each review as subjective or objective, we instead classified each *sentence* in each review as either "objective" or "subjective," keeping the probability being subjective $Pr_{subj}(s)$ for each sentence s. Hence, for each review, we have a "subjectivity" score for each of the sentences.

Based on the classification scores for the sentences in each review, we derived the average probability $AvgProb(r)$ of the review r being subjective defined as:

$$AvgProb(r) = \frac{1}{n}\sum_{i=1}^{n} Pr_{subj}(s_i) \quad (1)$$

where n is the number of sentences in review r and $s_1, \ldots, s_n$ are the sentences that appear in review r. Since the same review may be a mixture of objective and subjective sentences, we also kept of standard deviation $DevProb(r)$ of the subjectivity scores for each review, defined as:

$$DevProb(r) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\Big(Pr_{subj}(s_i) - AvgProb(r)\Big)^2} \quad (2)$$

Finally, to account for the cognitive cost required to read a review, we computed the average number of characters per sentence in the review, and the length of the review in sentences and in characters. Based on research in readability, these metrics are useful metrics for measuring how easy is for a user to read a review. For our study, we define the *Read* variable as the ratio of the length of the review in characters to the number of sentences.

4. ESTIMATING THE IMPACT OF REVIEW SUBJECTIVITY

Once we have derived the stylistic characteristics of each review, we can proceed to examine the economic impact of the subjectivity (or objectivity) of the review, after controlling for the other, easily observable numeric attributes. We ran two experiments that correspond to the two ranking schemes that we envision. The first experiment (Section 4.1) examines our techniques for measuring the effect of a review on product sales. Our results show how to rank the reviews for a merchant in terms of importance. Then, Section 4.2 presents our analysis on estimating the helpfulness of a review. Our preliminary results (Section 4.3) indicate how to rank a review for a consumer, even without the presence of peer review votes.

4.1 Effect of Subjectivity on Product Sales

We will first estimate the relationship between sales rank and subjectivity in reviews. We adopt a model similar to that used in [3] and [8], while incorporating measures for the quality of the content of the reviews. Chevalier and Mayzlin [3] and Forman, Ghose and Wiesenfeld [8] define the book's sales rank as a function of a book fixed effect and other factors that may impact the sales of a book. They also use a constant elasticity demand specification. The unit of observation in our analysis is a product-date, and the dependent variable is $\ln(SalesRank)$, the log of sales rank of product k in time t. Specifically, to study the impact of reviews and the quality of reviews on sales, we estimate the following model:

$$\begin{aligned}\ln(SalesRank)_{kt} = \alpha &+ \beta_1 \cdot \ln(RetailPrice_{kt}) + \\ &\beta_2 \cdot AvgProb_{k(t-1)} + \\ &\beta_3 \cdot DevProb_{k(t-1)} + \\ &\beta_4 \cdot Rating_{k(t-1)} + \\ &\beta_5 \cdot \ln(Reviews_{k(t-1)}) + \\ &\beta_6 \cdot \ln(Read_{k(t-1)}) + \\ &\beta_7 \cdot \ln(ElapsedDate_{kt}) + \mu_k + \varepsilon_{kt} \quad (3)\end{aligned}$$

where $AvgProb$, and $DevProb$ are variables that capture the degree of polarization or sentiment in reviews. μ_k is a product fixed effect that controls for unobserved heterogeneity across products. Note that increases in sales rank mean lower sales, so a negative coefficient increases sales. The control variables used include the retail price, the difference between the date of data collection and the release

[3] http://www.alias-i.com/lingpipe/

date of the product (*Elapsed Date*), the average numeric rating of the product (*Rating*), the number of reviews posted for that product (*Number of Reviews*), and the readability of the review (*Read*). We also used as control variables the minimum used price of the product, and the number of used goods available for sale. This did not affect the qualitative nature of the results and hence, they are omitted for brevity.

We also estimated a first-difference model with the dependent variable being $\delta((SaleRank_t) - (SalesRank_{t-1}))$. We estimated three different variations of the unit of time: at the daily level, weekly level and monthly level. The results were directionally similar to the ones presented above, and are omitted for brevity.

4.2 Effect of Subjectivity on Helpfulness

Consumers are more likely to post extreme reviews than more moderate reviews because highly positive or highly negative experiences with a product are more likely to motivate interpersonal communication behavior [7]. We use a well-known linear specification for our demand estimation [8]. Using the relationship in (1), we then estimate models of the form:

$$\begin{aligned}\ln(HELPFUL)_{kr} = \alpha + \beta_1(AvgProb)_{kr} + \\ \beta_2(DevProb)_{kr} + \\ \beta_3(MODERATE)_{kr} + \\ \beta_4 \ln(Read)_{kr} + \\ \beta_5 \ln(ElapsedDate)_{kr} + \mu_k + \varepsilon_{kr} \quad (4)\end{aligned}$$

where, k and r index product and review. The unit of observation in our analysis is a product-review and μ_k is a product fixed effect that controls for differences in the average helpfulness of reviews across products. The dependant variable *HELPFUL* is the log of the ratio of helpful votes to total votes received for a review.

4.3 Analysis

We find that an increase in the average subjectivity of a review leads to an increase in sales for audio —video players(see Table 3). It is statistically insignificant for digital cameras. It is possible that we may need a bigger data set to observe statistical significance, and our ongoing work is aimed at collecting additional data for that purpose. Products like electronic equipments have a number of attributes (or features) that consumers take into consideration while evaluating them. In such cases, more subjective reviews reduce the cognitive load of consumers and hence, this is more likely to be valued by users and results in higher sales.

The coefficient of *DevProb* has a negative relationship with sales rank in both categories although it is statistically significant only for digital cameras. In general this suggests that an increase in deviation leads to a decrease in sales rank, i.e., an increase in product sales. This means that reviews that have a mixture of objective, and highly subjective sentences have a positive effect on product sales, compared to reviews that tend to include only subjective or only objective information.

Using these results, it is now possible to generate a ranking scheme for presenting reviews to manufacturers of a product. The reviews that affect sales the most (either positively or negatively) are the reviews that should be presented first to the manufacturer. Such reviews tend to contain information that affects the perception of the customers for the product. Hence, the manufacturer can utilize such reviews, either by modifying future versions of the product or by modifying the existing marketing strategy (e.g., by emphasizing the good characteristics of the product). We should note that the reviews that affect sales most are not necessarily the same as the ones that customers find useful and are typically getting "spot-lighted" in review forums, like the one of Amazon. We present related evidence next.

With regard to the informativeness of reviews, our analysis reveals that for product categories such as audio and video equipments, and digital cameras, the extent of subjectivity in a review has a significant effect on the extent to which users perceive the review to be helpful. More interestingly, *DevProb* has always a positive relationship with helpfulness votes suggesting that consumers find more useful the reviews that have a wide range of subjectivity/objectivity scores across the sentences. In other words, reviews that have a mixture of sentences with objective and of sentences with extreme, subjective content are rated highly by users. This result is also corroborated by the sign of the coefficient on the *MODERATE* variable[4] on several of the product categories. The negative sign on this variable implies that as the review becomes more moderate or equivocal, it is considered less helpful by users.

This result is also in accordance with [8] who look at the numeric rating of reviews and assess its relationship with the percentage of helpfulness votes received by the review. Our analysis shows that we can estimate quickly the helpfulness of a review by performing an automatic stylistic analysis in terms of subjectivity. Hence, we can identify immediately reviews that have significant impact on sales and are expected to be helpful to the customers. Therefore, we can immediately rank these reviews higher and display them first to the customers. (This is similar to the "spotlight review" feature of Amazon which relies on the number of helpful votes posted for a review, and which has the unfortunate characteristic that requires a long time to pass before identifying a helpful review.)

5. VALIDATION WITH CONTENT ANALYSIS

In order to assess the validity of our automated content analysis using text mining techniques, we had two human coders do a content analysis on a sample of 1,000 reviews. The reviews were randomly chosen from across the seven product categories. The main aim was to analyze whether the review was informative and the extent to which it influenced a purchase decision. For this, the coders classified each review into categories based on whether the review influenced their decision to buy or not buy the product. Specifically, the coders had to answer two broad questions:

1. Is the review informative or not?
2. If you were interested in buying the product, would the review influence your decision?

For the first question, the potential answers were "yes" and "no" while for the second question, the coders could give one of the following four answers:

[4]The variable MODERATE is a dummy variable, taking values 0 or 1. We mark a review as *MODERATE* if its rating is 3 in the 5-star range.

Independent Variable	Audio-Video	Digital Camera
AvgProb	-1.47*** (.72)	1.27 (1.24)
DevProb	-0.69 (1.06)	-2.91*** (1.7)
Log (Amazon Price)	1.59*** (0.3)	6.2*** (0.61)
Log(Elapsed Date)	0.12 (0.07)	0.28** (0.13)
Average Rating	-0.016 (0.02)	-0.01 (0.03)
Log (Reviews)	0.6*** (0.15)	1.08*** (0.2)
Log(Read)	0.06(0.057)	-0.15* (0.08)
R^2	0.18	0.37

Table 2: The dependent variable is Log (Salesrank). Robust standard errors are listed in parenthesis; *, ** and * denote significance at 1%, 5% and 10%, respectively.**

Ind. Variable	Audio-Video	Digital Camera
AvgProb	-0.52* (.28)	-1.9*** (0.37)
DevProb	5.23***(0.42)	4.74*** (0.6)
Log(Elapsed Date)	0.044* (.02)	0.03 (.03)
MODERATE	-0.15*** (.03)	-0.06 (.04)
Log(Read)	0.21*** (.019)	0.21*** (.026)
R^2	0.07	0.1

Table 3: The dependent variable is Log (Helpful). Standard errors are listed in parenthesis; *, ** and * denote significance at 1%, 5% and 10%, respectively.**

1. Yes, positively
2. Yes, negatively
3. No, and
4. Uncertain

We measured the inter-rater agreement across the two coders, using the kappa statistic. The analysis showed a substantial agreement, with $\kappa = 0.739$. Similarly, we measured the agreement across the two raters for the second questions, using polychoric correlation and we found the agreement to be strong ($p < 0.05$). The results of the agreement tests, indicated that the reviews do exhibit common characteristics in terms of informativeness and in terms of influence.

Our next step was to identify the types of reviews that are considered useful by the users, and how this is reflected in the number of useful votes that they receive. Given the results of the annotation study, we wanted to identify the optimal threshold (in terms of percentage of helpful votes) that would separate the reviews that humans consider helpful from the non-helpful ones. We performed an ROC analysis, trying to balance the false positive rate and the false negative rate. Our analysis indicated that if we set the separation threshold at 0.6, then the error rates are minimized. In other words, if more than 60% of the votes indicate that the review is helpful, then we classify a review as "informative". Otherwise, the review is classified as "non-informative" and this decision achieves a good balance between false positive errors and false negative errors.

Of course, even if we have a good separation threshold, we still cannot say if a review is informative or not (and rank it properly) if we do not have votes from the peer reviewers. For this, we use our own subjectivity analysis technique, and we try to estimate the informativeness (or helpfulness) of a review, by using simply the text of the review. Towards addressing this, we first run our regressions, by removing from the data set the points that correspond to the reviews that our coders analyzed. We extracted the coefficients for the regressions, and then we examined whether the estimated coefficients can be used for prediction. For the 1,000 reviews in our manually annotated reviews, we used the regression coefficients (extracted during "training") to examine whether we can predict accurately the informativeness and influence of the review, by just using the text. Therefore, for each review, we could predict whether it is informative or not, and whether it is influential or not. We measured the accuracy of our predictions using the F-measure, which combines precision and recall into a single, concise metric. (The F-measure is the harmonic mean of precision and recall.) We present our estimates in Table 4.[5]

Product Category	F-measure
Audio-Video	0.85
Digital Camera	0.85
Overall	0.85

Table 4: F-Measure

In general, our results indicate that we can achieve good empirical performance. This means that we can derive from the text both the informativeness and the expected influence of each review. Overall, this means that once the review is submitted, we can rank it immediately without waiting for the peer reviews and the respective votes. Also, if a review is expected to have significant effect in the sales, the manufac-

[5]The F-measure across all categories pooled together was 0.74.

turer can identify it quickly and examine what attributes of the product are mentioned in the review, and are therefore important for marketing purposes.

6. RELATED WORK

Our research program is inspired by previous studies about opinion strength analysis. While prior work in computer science has extensively analyzed and classified sentiments in online opinions [12, 14, 15, 17, 21, 23], they have not examined the economic impact of the reviews. Similarly, while prior work has looked at how the average rating of a review or social factors (such as self-descriptive information) is related to the proportion of helpful votes received by a review, it has not looked at how the textual sentiment of a review affects it. Similarly, prior work has shown that the volume and valence of online product reviews influences product sales such as books and movies [3, 7, 8] but this stream of research did not account for the textual content in those reviews while estimating their impact on sales. To the best of our knowledge no prior work has combined sentiment analysis techniques from opinion mining with economic methods to evaluate how the content of reviews impacts sales. Our research papers aim to make a contribution by bridging these two streams of work.

We also add to an emerging stream of literature that combines economic methods with text mining [5, 9, 16]. For example, Das and Chen [5] examined bulletin boards on Yahoo! Finance to extract the sentiment of individual investors about tech companies and about the tech sector in general. They have shown that the aggregate tech sector sentiment predicts well the stock index movement, even though the sentiment cannot predict well the individual stock movements. There has also been related work on studying connections between online content such as blogs, bulletin boards and consumer reviews, and consumer behavior, in particular purchase decisions. Gruhl et al. [10] analyzed the correlation between online mentions of a product and sales of that product. Using sales rank information for more than 2,000 books from Amazon.com, Gruhl et al. demonstrated that, even though sales rank motion might be difficult to predict in general, online chatter can be used to successfully predict *spikes* in the sales rank.

7. CONCLUSIONS

We contribute to previous research that has explored the informational influence of consumer reviews on economic behavior such as how online reviews increase sales and the impact of critics reviews on box office revenues by suggesting that patterns of sentiment may influence purchasing decisions over and above the numeric ratings that online consumer reviews display. The present paper is unique in looking at how sentiment in text of a review affects product sales and the extent to which these reviews are informative as gauged by the affect of sentiments on helpfulness of these reviews. We also find that reviews which tend to include a mixture of subjective and objective elements are considered more informative (or helpful) by the users. However, for the effect on sales, we need to conduct further investigations and potentially examine the interactions of the subjectivity metrics with the numeric rating of the review. In terms of subjectivity and effect on helpfulness, we observe that for feature-based goods, such as electronics, users prefer reviews to contain mainly objective information with a few subjective sentences. In other words, the users want the reviews to mainly confirm the validity of the product description, giving a small number of comments (not giving comments decreases the usefulness of the review). For experience goods, such as movies, users prefer a brief description of the objective elements of the good (e.g., the plot) and then the users expect to see a personalized, highly sentimental positioning, describing aspects of the good that are not captured by the product description. Based on our findings, we can identify quickly reviews that are expected to be helpful to the users, and display them first, improving significantly the usefulness of the reviewing mechanism to the users of the electronic marketplace. We are collecting additional data to enhance the scope of our findings.

Acknowledgments

We thank Rhong Zheng for assistance in data collection. This work was partially supported by a Microsoft Live Labs Search Award, a Microsoft Virtual Earth Award, a New York University Research Challenge Fund grant N-6011, and by NSF grants IIS-0643847 and IIS-0643846. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the Microsoft Corporation or of the National Science Foundation.

References

[1] Archak, N., Ghose, A., and Ipeirotis, P. G. Show me the money! Deriving the pricing power of product features by mining consumer reviews. In *Proceedings of the Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2007)* (2007).

[2] Berndt, E. R. *The Practice of Econometrics: Classic and Contemporary*. Addison-Wesley, 1996.

[3] Chevalier, J. A., and Mayzlin, D. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research 43*, 3 (Aug. 2006), 345–354.

[4] Cui, H., Mittal, V., and Datar, M. Comparative experiments on sentiment classification for online product reviews. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-2006)* (2006).

[5] Das, S. R., and Chen, M. Yahoo! for Amazon: Sentiment extraction from small talk on the web. Working Paper, Santa Clara University. Available at http://scumis.scu.edu/~srdas/chat.pdf, 2006.

[6] Dave, K., Lawrence, S., and Pennock, D. M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International World Wide Web Conference (WWW12)* (2003), pp. 519–528.

[7] Dellarocas, C., Awady, N. F., and Zhangz, X. M. Exploring the value of online product ratings in revenue forecasting: The case of motion pictures. Working Paper, Robert H. Smith School Research Paper, 2007.

[8] Forman, C., Ghose, A., and Wiesenfeld, B. Examining the relationship between reviews and sales: The role of social information in electronic markets. Working Paper CeDER-06-09, Stern School of Business, New York University. Available at http://hdl.handle.net/2451/14809, 2006.

[9] Ghose, A., Ipeirotis, P. G., and Sundararajan, A. Opinion mining using econometrics: A case study on reputation systems. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL 2007)* (2007).

[10] Gruhl, D., Guha, R., Kumar, R., Novak, J., and Tomkins, A. The predictive power of online chatter. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2005)* (2005), pp. 78–87.

[11] Hatzivassiloglou, V., and McKeown, K. R. Predicting the semantic orientation of adjectives. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'97)* (1997), pp. 174–181.

[12] Hu, M., and Liu, B. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004)* (2004), pp. 168–177.

[13] Hu, N., Pavlou, P. A., and Zhang, J. 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 (EC'06)* (2006), pp. 324–330.

[14] Kim, S.-M., and Hovy, E. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)* (2004), pp. 1367–1373.

[15] Lee, T. Y., and Bradlow, E. T. Automatic construction of conjoint attributes and levels from online customer reviews. University of Pennsylvania, The Wharton School Working Paper OPIM WP 06-08-01, 2006.

[16] Lewitt, S., and Syverson, C. Market distortions when agents are better informed: The value of information in real estate transactions. Working Paper, University of Chicago, 2005.

[17] Liu, B., Hu, M., and Cheng, J. Opinion observer: Analyzing and comparing opinions on the Web. In *Proceedings of the 14th International World Wide Web Conference (WWW 2005)* (2005), pp. 342–351.

[18] Nigam, K., and Hurst, M. Towards a robust metric of opinion. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text* (2004), pp. 598–603.

[19] Pang, B., and Lee, L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)* (2004), pp. 271–278.

[20] Pang, B., and Lee, L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)* (2005).

[21] Pang, B., Lee, L., and Vaithyanathan, S. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)* (2002).

[22] Snyder, B., and Barzilay, R. Multiple aspect ranking using the good grief algorithm. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL 2007)* (2007).

[23] Turney, P. D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)* (2002), pp. 417–424.

[24] Wilson, T., Wiebe, J., and Hwa, R. Recognizing strong and weak opinion clauses. *Computational Intelligence 22*, 2 (May 2006), 73–99.