Designing Ranking Systems for Consumer Reviews:
The Impact of Review Subjectivity on Product Sales and Review Quality

Anindya Ghose, Panagiotis G. Ipeirotis
{aghose, panos}@stern.nyu.edu

Department of Information, Operations, and Management Sciences
Stern School of Business
New York University
44 West 4th Street, New York, NY-10012

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 wants 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 in general, 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.

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. 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. Reviews are either allotted an extremely high rating or an extremely low rating. In such situations, the average numerical star rating may not convey a lot of information to a prospective buyer, since the reader has to read the 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,

This work has been partially supported by a Microsoft Live Labs Search Award

where customers give "helpful" votes to other reviews. 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 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. However, 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. Hence, as an important contribution of this paper 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 (Pang, Lee and Vaithyanathan 2002, Turney 2002, Hu and Liu 2004, Kim and Hovy 2004, Liu, Hu and Cheng 2005), and explored how *automatic* procedures can be used for obtaining conjoint attributes and levels through the use of natural language processing, statistical clustering methods (Lee and Bradlow 2006), they have not examined their economic impact. Similarly, prior work has shown that the volume and valence of online product reviews influences product sales such as books and movies (Dellarocas et. al. 2005, Forman, Ghose and Wiesenfeld 2006, Chevalier and Mayzlin 2006) 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.

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. Finally, Section 5 provides some additional discussion and concludes the paper.

2. Data

To conduct our study, we created a panel data set 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 September 2005-April 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, and digital cameras. For 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. We recorded the numerical rating and we constructed a dummy variable to differentiate between extreme reviews which are unequivocal and moderate reviews. Specifically, ratings in the middle of the scale (3) was classified as

MODERATE while ratings at either of the two endpoints of the scale (1, 2, 4, or 5) were classified as *EXTREME.*[1]

3. Text Mining for Objectivity and Subjectivity Estimation

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. In our setting, the numerical rating score already gives the polarity of the review, 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 relied on existing literature of subjectivity estimation from computational linguistics (Pang and Lee, 2004). 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. We trained our classifier using as "objective" documents the product descriptions of each of the 1,000 products in our data set, and we retrieved randomly reviews to construct the "subjective" part of the training set. 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.

After constructing the classifiers, we classified each sentence in each review as either "objective" or "subjective", keeping the confidence score for each classification. Hence, for each review, we have a "subjectivity" score for each of the sentences. Based on the classification scores, we derived the average probability of the review being subjective. We denote this probability by ***AVGPROB***. Since the same review may be a mixture of objective and subjective sentences, we also kept of deviation of the sentence scores for each review, denoted by ***DEVPROB***. 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. Empirical Econometric Analysis and Results

Once, we have derived the stylistic characteristics of each review, we 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, described in 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 results indicate how to rank a review for a consumer, even without the presence of peer review votes.

[1] Our results are not sensitive to the specific coding method. In particular, we used an alternate specification in which ratings between and inclusive of 2 and 4 were classified as *MODERATE* while ratings at either of the two endpoints of the scale (1, or 5) were classified as *EXTREME*. This does not change the qualitative nature of the results.

4.1 Effect of Subjectivity on Product Sales

We first estimate the relationship between sales rank and subjectivity in reviews. We adopt the same *difference-in-difference* strategy used in Chevalier and Mayzlin (2006), while incorporating measures for the quality of the content of the reviews. Chevalier and Mayzlin (2006) 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 *log(SalesRank)*, the log of sales rank of product *i* in time *t*. Specifically, to study the impact of reviews and the quality of reviews on sales, we estimate the following model:[2]

$$log(SALESRANK_{it}) = \alpha + \beta_1 AMAZONPRICE_{it} + \beta_2 log(ELAPSEDDATE_{it}) + \beta_3 AVGPROB_{it} + \beta_4 DEVPROB_{it} + \mu_i + \varepsilon_{it}$$

where *AVGPROB* , and *DEVPROB* are variables that capture the degree of polarization or sentiment in reviews and μ_i 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 Amazon price *Amazon Price*, the difference between the date of data collection and the release date of the product *Elapsed Date,* the average numeric rating of the product, *Average Rating,* the number of reviews posted for that product, *Number of Reviews,* and the readability of the review, *READ*.[3]

Independent Variable	Audio-Video	Digital Camera	Computer	DVD	Video Game
AVGPROB	-1.47*** (.72)	1.27(1.24)	-1.64** (0.6)	-1.54** (0.7)	0.75 (1.01)
DEVPROB	-0.69(1.06)	-2.91***(1.7)	-1.3 (0.98)	-0.15(1.4)	3.57***(1.3)
Log (Amazon Price)	1.59***(0.3)	6.2*** (0.61)	3.8*** (0.6)	1.61*** (0.2)	1.09*** (.13)
Log(Elapsed Date)	0.12 (0.07)	0.28**(0.13)	0.85***(0.1)	0.7*** (0.03)	-0.2 (0.12)
Average Rating	-0.016(0.02)	-0.01(0.03)	-0.01 (0.01)	.012 (0.02)	-0.05* (.027)
Log(Number of Reviews)	0.6*** (0.15)	1.08***(0.2)	3.76*** (0.2)	-0.3***(0.08)	0.94***(0.3)
Log(Read)	0.06(0.057)	-0.15*(0.08)	-0.1** (0.05)	-0.07 (0.05)	0.11**(0.06)
R^2	0.18	0.37	0.3	0.38	0.37

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

We find that an increase in the average subjectivity of a review leads to an increase in sales for audio - video players, computers and for DVDs. It is statistically insignificant for digital cameras, video games and software. 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.

[2] Note that prior work in this domain has generally transformed the dependent variable (sales rank) into quantities using the specification similar to Ghose, Smith and Telang (2006) since that paper needed to estimate demand. However, in our paper we are not interested in estimating demand, and hence we do not need to make the transformation.

[3] 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.

In general, the coefficient of *DEVPROB* has a negative relationship with sales rank suggesting 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. However, we have some results (i.e., in VideoGames) that contradict our hypothesis. This suggests that we need to examine further this correlation, perhaps by examining the effect of *DEVPROB* in positive, negative, and moderate reviews.

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 "spotlighted" in review forums, like the one of Amazon. We present related evidence next.

4.2 Effect of Subjectivity on Helpfulness of Reviews

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 (Dellarocas et al., 2005). We use a well-known linear specification for our demand estimation. Using the relationship in (1), we then estimate models of the form:

$$log(HELPFUL_{it})=\alpha_0+\alpha_1(AVGPROB_{it})+\alpha_2(DEVPROB_{it})+\alpha_3 log(ELAPSEDDATE_{is})+\Omega' X+\mu_i+\varepsilon_{it} \quad (2)$$

where, i and t index product and time. The unit of observation in our analysis is a product-review and μ_i is a product fixed effect that controls for differences in the average helpfulness of reviews across books. The dependant variable *HELPFUL* is the ratio of helpful votes to total votes received for a review.

Our analysis reveals that for product categories such as audio and video equipments, digital cameras, and computers 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. The negative sign implies that as the review becomes more moderate or equivocal, it is considered less helpful by users.

Independent Variable	**Audio-Video**	**Digital Cam-**	**Computer**	**DVD**	**Video Game**
AVGPROB	0.58*** (.19)	-1.64*** (0.4)	0.92*** (0.3)	-0.4 (1.34)	-0.42 (1.29)
DEVPROB	4.11*** (0.5)	4.93*** (0.6)	2.7*** (0.5)	6.9** (2.9)	8.7*** (2.0)
Log(Elapsed Date)	0.017(.03)	0 .09*** (.03)	-0.03 (.04)	-0.12 (.42)	-0.012(.17)
MODERATE	-0.1*** (.04)	-0.065 (.04)	-0.06** (.02)	-0.07 (0.17)	-0.11 (0.15)
Log(NumberofReviews)	-0.07 (.06)	-0.02 (.054)	.01(.065)	.11 (0.31)	.1 (0.37)
Log(READ)	0.2*** (.024)	0.22*** (.03)	0.17** (.02)	0.24** (0.1)	0.22** (0.08)
R^2	0.12	0.11	0.06	0.06	0.07

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

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 *immedi-*

ately 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. Conclusion

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.

References

1. Chevalier, J., and D. Mayzlin. 2006. The effect of word of mouth online: Online book reviews. *Journal of Marketing Research,* forthcoming.
2. Dellarocas, C., N. Awad, and M. Zhang. 2005. Using online ratings as a proxy of word-of-mouth in motion picture revenue forecasting. Working Paper, University of Maryland.
3. Forman. C, A. Ghose, and B. Wiesenfeld. 2006. A Multi-Level Examination of the Impact of Social Identities on Economic Transactions in Electronic Markets. NYU CeDER Working Paper # 06-09.
4. Ghose, A., M. Smith, and R. Telang. 2006. Internet exchanges for used books: An empirical analysis of product cannibalization and welfare impact. *Information Systems Research* 17(1): 3–19.
5. Hu, M., and B. Liu. 2004. Mining and Summarizing Customer Reviews. *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*
6. Kim, S., and E. Hovy. 2004. Determining the Sentiment of Opinions. *Proceedings of COLING-04.* pp. 1367-1373. Geneva, Switzerland.
7. Lee, T. and E. Bradlow. 2006. Automatic Construction of Conjoint Attributes and Levels from Online Customer Reviews. Working Paper, Wharton School.
8. Pang, B., Lee, L., and S. Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of Empirical Methods in Natural Language Processing*.
9. Pang, B., and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. . *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics.*
10. Turney, P. D. 2002. Thumbs Up Or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.*