

Stay Out of My Forum! Evaluating Firm Involvement in Online Ratings Communities

Neveen Awad and Hila Etzion

1. INTRODUCTION

A growing number of online retailers are enabling and encouraging consumers to post reviews of the products sold on their website. These customer reviews often consist of a numeric value (rating), along with some optional text describing the reviewer's experience with the product. Recent research suggests that consumers utilize online ratings in making their purchase decisions (Chevalier and Mayzlin, 2003; Senecal and Nantel, 2003). However, there is controversy related to the reliability of online reviews as well as to how well they reflect the opinions of the population of consumers. Anecdotal evidence suggests that some of this information may be biased and is sometimes provided anonymously by the product companies themselves (White 1999; Harmon 2004). Some firms have attempted to address inherent biases, with the hope of increasing sales, by filtering customer reviews. However, to date there is very little research regarding whether online retailers should filter customer reviews, and when such involvement will result in a business advantage.

In this paper we examine retailer's involvement with online word of mouth posted on its website by studying, analytically and empirically, the relationship between posted online reviews and sales. Empirically, we examine the relationship of online ratings with purchase transactions at a large online retailer before and after it changed its policy for filtering reviews. Specifically, we examine the impact of online ratings on sales across two different filtering strategies. Before March 3, 2005, the firm's filtering strategy was to filter out all reviews that reflected negatively on the firm or their products in any way. Thus the filtering was done by the marketing department, and the goal was to only keep reviews that would enhance sales. After March 3, 2005, the filtering was given to the online experience department. The strategy of filtering changed to one of "noise reduction". Thus, reviews were filtered out only if they were deemed to provide no value (positive or negative). As such, comments including profanities, or comments not having to do with the product were filtered.

Our study provides affirmative answers to several important questions: *Are online reviews correlated with online transactions?* Our results provide evidence for the claim that online ratings are associated with online purchases. *Does filtering of online reviews affect the impact of these reviews on online transactions?* Online retailers have a wide range of approaches to filtering customer product reviews, with many struggling to find the correct balance. The retailer we study in this paper changed its filtering strategy and filtering team in March of 2005, which allowed us to empirically compare the relationship of reviews with sales across two filtering strategies: a strategy that filters out most negative reviews, and a strategy that filters only noise. We find that a firm's filtering strategy impacts the relationship between average rating, number of extreme ratings (positive or negative) and sales.

To address the last question thoroughly, and examine whether it is optimal for the retailer to filter bad reviews, we analytically model an online retailer that sells two competing products. The products are imperfect substitutes, and therefore the demand for a product depends not only on its price and ratings, but also on the substitute's price and ratings. We find that if the seller compares only two strategies: filtering bad reviews and not filtering bad reviews across both products, then not filtering dominates when the proportion of bad reviews for the more profitable product is small enough relative to the proportion of bad reviews for the less profitable product.

The rest of the paper is organized as follows. Section 2 presents the data set. Sections 3 and 4 describe our empirical methodology and present the empirical results. In Section 5 we present the analytical model. In Section 6 we discuss the implications of the empirical and analytical findings, conclude, and describe the next steps of this work.

2. DATA SET

Our data for this study consists of individual product characteristics, purchase transactions, and user reviews. These data were collected from a large online retailer, and the dates of the data range from April 16th, 1999 to February 2nd, 2006. The firm changed its reviews filtering method on March 3rd 2005.

The user reviews data consisted of an integer numerical rating that ranged from 5 (best) to 1 (worst) and an optional text review of the product. Before these ratings are published on the online retailer's website, they are first put in a queue to be assessed by the review filtering team. As the team goes through the reviews, they either approve or reject the review. We restrict our analysis to only the approved reviews, since those are the ones visible to consumers. Table 1 presents summary statistics for the two periods: before and after the filtering strategy changed.

Table 1 – Summary Statistics

	Average across products					
	Rating	Variance of Rating	Number of Reviews	Number of 1s	Number of 5s	Sales
4/16/99 - 3/03/05	4.662	0.04	1.838	0.05	1.428	1.101
3/04/05- 2/02/06	4.573	0.055	2.641	0.101	1.942	1.078
% Change	-2%	38%	44%	102%	36%	-2%

Table 1 shows an increase of 102% in the average number of 1's per product. In addition the average rating went down by 2%, reflecting the fact that the retailer stopped filtering reviews with low rating. We also see that though the average number of 5's per product increased by 36%; this increase is much smaller than the increase in the average number of 1's. It is not surprising that sales went down, because in the first period we consider sales from 5 years while in the second period we consider sales from only 10 months.

We run two separate empirical models: 1) using 11 months of sales data from before the change in the filtering strategy along with review data posted between 4/16/99 and 3/03/05; and 2) using 11 months of sales data from after the filtering change along with review data posted between 4/16/99 and 2/02/06. We limit the “before” sales period to 11 month to control for the difference in time frame of available data before and after the strategy change. We also tested the model using the full sales data set before the change, which encompasses just under 6 years of data. The results for this last model are similar in sign and significance to the results of the before model that used only 11 months of data.

3. EMPIRICAL METHODOLOGY

The goal of this study is to examine the relationship between purchase activities and the review information, before and after the firm changed the review filtering strategy. The dependent variable is the total amount spent per product¹. The independent variables include: *volume*, measured as total number of reviews per product (Godes and Mayzlin, 2002), *valence*, measured as average rating per product, and *density*, measured as number of reviews per product divided by number of transactions per product (Dellarocas et al, 2004), . The theory behind volume is that the more consumers discuss a product, the higher the chance that other consumers will become aware of it. Density takes this theory one step further by normalizing the number of reviews of a product by the number of transactions. The theory behind valence is that the average rating is a proxy for the quality of the product. Our independent variables also include: *number of extreme negative or extreme positive* (e.g. the number of reviews that were the worst rating (1 on a 1 to 5 scale), or the best rating (5 on a 1 to 5 scale)) (Chevalier and Mayzlin, 2003), and *variance of the ratings per product*, which approximates the range of disagreement between reviewers. To assess the impact of products which are potential complements, we also include *category valence*, which is the average rating of products classified within the same sub-category.

Beyond the ratings given by the reviewers, there might be additional information contained in the review's text. Coding the actual text in the reviews has been shown to produce rather noisy results (Godes and Mayzlin 2003). Thus, in accordance with prior research, we include just *review length*, measured as number of words; Prior literature has suggested that a longer review may suggest a more “mixed” review (Chevalier and Mayzlin, 2003). Lastly, we categorize the products into four groups of prices (\$0-\$20, \$20-\$50, \$50-\$150, \$150-\$10,000) using three dummy variables

Consider a product i that is sold on the retailer's website. The product i belongs to the category j and the department k . Purchase activities are set up as the following:

$$\begin{aligned}
 spending_{i(j(k))} = & \beta_0 - i(j(k)) + \beta_1 Volume + \beta_2 Valence + \beta_3 Category_Valence + \beta_4 Variance + \beta_5 num_rev_1 \\
 & + \beta_6 num_rev_5 + \beta_7 ReviewLength + \beta_8 ProductPrice + \beta_9 Density + \beta_{10}(price_category_dummy) + \varepsilon_{i(j(k))}
 \end{aligned} \quad (1)$$

Where $\varepsilon_{i(j(k))}$ is the random error for product i . We assume that the coefficients of the average ratings vary with the standard deviation of the corresponding rating, because the average rating is less informative when there is more variation in the ratings. That is, we specify $\beta_1 = \beta_1' + \gamma_1(std_dev_t)$. In addition, some unobservable characteristics across categories and departments are hypothesized to randomly influence the intercept $\beta_0 - ijk$. That is, $\beta_0 - i(j(k)) = \beta_0 - i(j) + \varepsilon_k$, and $\beta_{0i(j)} = \beta_0 - i + \varepsilon_{j(k)}$. Thus, $\beta_0 - i(j(k)) = \beta_0 - i + \varepsilon_k + \varepsilon_{j(k)}$. After these adjustments, we estimate the function:

$$\begin{aligned}
 spending_{i(j(k))} = & \beta_0 - i(j(k)) + \beta_1 Volume + \beta_2 Valence + \gamma_2(\sqrt{Variance}) * Valence + \beta_3 Category_Valence + \gamma_3(\sqrt{Category_Variance}) * Category_Valence \\
 & + \beta_4 Variance + \beta_5 num_rev_1 + \beta_6 num_rev_5 + \beta_7 ReviewLength + \beta_8 ProductPrice + \beta_9 Density + \beta_{10}(price_category_dummy) + \varepsilon_{i(j(k))} + \varepsilon_{j(k)} + \varepsilon_k
 \end{aligned} \quad (2)$$

¹ We also used number of transaction per product as the dependent variable, but due to space limitations, we do not include those results in this draft.

As described above, we run this model with 11 months of sales data before the change in filtering strategy and 11 months of sales data after the change in filtering strategy.

4. EMPIRICAL RESULTS

The estimation results for the above specified model are as follows:

Table 2 – Purchase amount before and after the filtering strategy change

Effect	Negative Review Filtering (Before)		Noise Reduction Filtering (After)	
	Before (β)	t value (Pr> t)	After (β)	t value (Pr> t)
β_0	-29.624	-1.38 (0.216)	11.676	0.4 (0.692)
Volume	2.753*	2.12 (0.034)	-0.784	-1.24 (0.216)
Valence	4.451**	2.81 (0.005)	-2.484*	-2.45 (0.014)
Category_Valence	7.098*	1.76 (0.083)	6.303	0.99 (0.320)
Variance	-0.997	-0.79 (0.431)	-2.634**	-2.99 (0.003)
review_n_1	13.224***	3.75 (0.0002)	-2.828*	2.00 (0.054)
review_n_5	-3.681**	-2.26 (0.024)	1.117*	1.88 (0.068)
Review_length	-0.036	-0.46 (0.642)	0.021	0.37 (0.713)
Product_Price	0.350***	120.42 (<0.0001)	0.331***	150.76 (<0.0001)
Density	-4.385**	-2.85 (0.004)	-3.297*	-1.94 (0.052)
dummy_1	3.164	0.93 (0.351)	3.888	1.61 (0.108)
dummy_2	9.352*	2.36 (0.019)	9.392*	3.11 (0.002)
dummy_3	14.345**	3.26 (0.001)	14.842***	4.26 (<0.0001)

***: p < 0.001, **: p < 0.01, *: p < 0.10

There are two main differences that we see in the results for before and after the change in filtering strategy. Before the change the number of '1' ratings is positive and significantly associated with the purchase amount, whereas the number of '5' ratings is negative and significantly associated with the purchase amount. We expect that this somewhat counterintuitive result is due to their being a small proportion of 1's when the marketing department was in charge of the filtering, therefore the presence of 1's actually added credibility to the valence score, and therefore was associated with an increase in purchasing. Similarly, we expect that the negative correlation between the number of 5's and the purchase amount is due to the perception of review bias increasing with the greater number of extreme positive reviews. As a result, consumers were likely to suspect of the reviews, and therefore purchase amount was lower for products which appeared to have more bias in their reviews. Notice that *valence* is positive and significantly associated with sales before the filtering change. That is, although extreme rating values have the opposite effect than one would expect, sales still increase with the average rating. This supports our argument that the presence of a small number of 1's gave validity to a product's ratings.

In contrast, we see that after the filtering strategy changed to one of noise reduction the impact of the extreme value reviews reversed: the number of '1' ratings is negative and significantly associated with the purchase amount, whereas the number of '5' ratings is positive and significantly associated with the purchase amount. After the change, when the firm allowed negative reviews to surface, consumers seem to perceive the review information provided as accurate, rather than bias. Therefore, since consumers perceive the reviews to be revealing true information, negative reviews are having the expected negative effect, and positive reviews are having the expected positive effect. In addition, when the extreme value reviews have their expected effect, the valence and variance are negative and significantly associated with the purchase amount. A greater variance indicates more disagreement among the posted ratings. Thus consumers likely perceived greater variance as greater risk, resulting in a negative association with purchase amount. The negative association of the valence score after the filtering change is a very interesting result; we suspect that in the presence of greater variance of scores (extreme scores of both types are being displayed) consumers are focusing more on the extreme value ratings than on the valence. Thus even if the average rating score is lower, if there are more '5' reviews, and less '1' reviews, the consumer will be more apt to purchase that product. We further investigate this potential switch in consumer assessment through our analytical model. Our empirical results show a significant positive effect of *category_valence* on sales when the firm was filtering negative reviews (before). This result suggests that average ratings of compliments or substitutes have effect on sales of a product. The analytical model further investigates the impact of similar products that are imperfect substitutes.

5. ANALYTICAL MODEL

An online retailer offers two competing products. Retail firms sell products produced by other companies, and often can not control the products' posted prices. However, retailers can control which reviews are being displayed on their website. That is, they can filter the reviews submitted by consumers. To simplify the model, and to focus on the affect of the filtering

strategy, we assume that consumers who have purchased a product can submit one of two ratings: a good rating, G , and a bad rating, B . A bad rating has a score of -1 and a good rating has a score of 1. Table 3 summarizes the notation.

G_i	the number of good ratings for product i
B_i	the number of bad ratings for product i
N_i	the total number of reviews for product i
p_i	price of product i to consumers
c_i	cost of product i (for the retailer)
S_i	average rating for product i
$\pi(x,y)$	the seller's profit when there are x (y) bad reviews for product 1 (2)

Table 3: Notation

The retailer can filter the reviews. We assume that the retailer uses the same filtering strategy for both products (though, as shown later, this might not be optimal) and compare the profit from using the following two strategies: 1) Strategy F : filter/don't show bad reviews; and 2) Strategy NF : Show all reviews.

The products are *imperfect substitutes* and the demand functions are given by:

$$\begin{aligned} D_1(B_1, B_2) &= A_1 - bp_1 + dp_2 + \alpha S_1 - \delta S_2 \\ D_2(B_1, B_2) &= A_2 - bp_2 + dp_1 + \alpha S_2 - \delta S_1 \end{aligned} \quad (3)$$

Where $S_i=(G_i-B_i)/(G_i+B_i)$ is the average rating (score) for product i . The first three terms in the expression for D_i give the demand for the product in a market with no feedback system. We assume that $b>d$, so that a product's own price effect is stronger than the cross price effect (if the two prices go up by the same amount, the demand for both products goes down). The last two terms in the expression for D_i capture the effect of the difference in the average rating. The maximum affect the feedback system might have on the demand for a product is given by $(\alpha+\delta)$. That is, due to the feedback system, sales of product i (j) can increase (decrease) by at most $\alpha+\delta$ units, which happens when all of product i 's ratings are good and all of product j 's ratings are bad.

An increase in the demand of product i can be attributed to an increase in its perceived value (due to a good average rating). The positive valence of a product attracts consumers who initially were not going to buy any of the two products, as well as consumers who initially (in the absence of a feedback system) would have preferred product j .

If $\alpha=\delta$, then when the two products have the same average rating, i.e. $S_i=S_j=S$, the demand for each product is the same as when there is no feedbacks at all. Here, the feedback system only transfers demand from one product to the other, but cannot increase or decrease total demand. What one product loses in sales is the other product's gain. On the other hand, If $\alpha>\delta$, then when the two products have the same average rating, the demand for each product differs from the demand in a market with no feedback system by $\pm(\alpha-\delta)S$. In addition, if the average rating for the two products increases (decreases) by the same amount, the demand for both products goes up (down), and thus total sales can be higher or lower than the total sales in a market with no feedback system. The more dissimilar the products are, the smaller the cross affect of ratings should be, i.e. δ decreases. The seller's profit is given by

$$\pi(B_1, B_2)=(p_1-c_1)D_1+(p_2-c_2)D_2. \quad (4)$$

We compare the profit of the seller under the two filtering strategies described above, F and NF , for a given profile of submitted reviews (G_1, B_1, G_2, B_2) and products margins, (p_2-c_2) and (p_1-c_1) .

Lemma 1 The retailer's profit increases as the number of bad reviews for product i decreases if and only the ratio of the margins (margin for product j divided by margin for product i) is smaller than α/δ .

Proof. $\partial\pi(B_i, B_j)/\partial B_i < 0$ if and only if $(p_j - c_j)/(p_i - c_i) < \alpha/\delta$. (5)

According to Lemma 1, if $\delta=\alpha$ (ratings can only transfer demand between the products) the retailer should filter bad reviews only for the product with the higher margin. If $\delta=0$ (ratings do not transfer demand between products) the retailer should filter bad reviews for both products. And finally, if $\delta<\alpha$ (i.e. $\alpha/\delta>1$) it is profitable to filter bad reviews **at least** for the product with the higher margin (because when $(p_j-c_j)/(p_i-c_i)>1$, that is j is the product with the higher margin, then necessarily $(p_i-c_i)/(p_j-c_j)<1$). Notice that as δ decreases it becomes optimal to filter bad reviews for both products.

Though Lemma 1 gives the optimal filtering strategy when the seller can filter reviews selectively only for one of the products, such a strategy might not be possible due to the manufacturers' response, or might be difficult to execute when prices of the products change frequently. Hence, considering only the two extreme (and homogenous) strategies, F and NF , described above, the first dominates the latter if and only if $\Delta=\pi(0,0)-\pi(B_1, B_2)>0$, where:

$$\Delta = 2(B_1 N_2 (\alpha(p_1 - c_1) - \delta(p_2 - c_2)) + B_2 N_1 (\alpha(p_2 - c_2) - \delta(p_1 - c_1)))/(N_1 N_2). \quad (6)$$

Clearly if Inequality (5) holds for both products- that is if it is optimal to filter bad reviews for each of the products when considered separately, then strategy F dominates strategy NF . If inequality (5) holds for *only* one of the products, it would hold for the product with the higher margin. That is if $(p_j - c_j) > (p_i - c_i)$ we would have: $(p_j - c_j)/(p_i - c_i) > \alpha/\delta$ and $(p_i - c_i)/(p_j - c_j) < \alpha/\delta$. Then strategy NF dominates strategy F if and only if the ratio $(B_i/N_i)/(B_j/N_j)$, where j is the product with the higher margin, is larger than a threshold value. Proposition 1 summarizes our findings.

Proposition 1. NF dominates F if and only if $(p_j - c_j)/(p_i - c_i) > \alpha/\delta$, where i is the product with the lower margin, and

$$\frac{B_i / N_i}{B_j / N_j} > \frac{\alpha(p_j - c_j) - \delta(p_i - c_i)}{\delta(p_j - c_j) - \alpha(p_i - c_i)}. \quad (7)$$

The RHS in (7) is positive when $\alpha/\delta > 1$ because $(p_j - c_j)/(p_i - c_i) > \alpha/\delta$ and $(p_j - c_j) > (p_i - c_i)$. According to Proposition 1, if the product with the higher margin, in this case product j , has a small proportion of bad reviews when compared with the product with the lower margin, then the seller is better off not filtering at all to filtering for both products. If $\delta = \alpha$ (the products are close to being perfect substitutes), NF dominates F if and only if $(B_i/N_i) > (B_j/N_j)$. If the products have the same profit margin and $\alpha/\delta > 1$, then F always dominates NF .

However, what if by filtering all bad reviews the retailer actually changes the structure of the demand? Next we assume that if consumers do not observe any bad reviews, then instead of using the average rating, they compare the number of good reviews across the competing products. In such a case the demand functions are given by:

$$\begin{aligned} D_1(B_1, B_2) &= A_1 - bp_1 + dp_2 + \beta(\alpha S_1 - \delta S_2) + (1 - \beta)(\alpha AS_1 - \delta AS_2) \\ D_2(B_1, B_2) &= A_2 - bp_2 + dp_1 + \beta(\alpha S_2 - \delta S_1) + (1 - \beta)(\alpha AS_2 - \delta AS_1) \end{aligned} \quad (8)$$

Where $\beta=1$ if $B_1 + B_2 > 0$ and $\beta=0$ otherwise, and $AS_i = G_i/(G_i + G_j)$ is the Adjusted Score for product i .

Alternatively, we can examine the optimal filtering strategy assuming consumers suspect bias for a product that has only positive reviews, even if other products have bad reviews. In this case

$$D_i(B_i, B_j) = A_i - bp_i + dp_j + \alpha(\beta_i S_i + (1 - \beta_i) AS_i) - \delta(\beta_j S_j + (1 - \beta_j) AS_j) \quad (9)$$

Where $\beta_i=1$ if $B_i > 0$ and $\beta_i=0$ otherwise.

Given that consumers adjust their use of the feedback information based on whether they perceive it to be bias or not, is it still optimal for the retailer to filter bad reviews for the product with the higher margin? Under what conditions filtering bad reviews for all products dominate not filtering for both? This would be the next step in our analysis

6. DISCUSSION AND CONCLUSION

Our initial results suggest several important issues. First, our empirical results show positive support that online ratings are significantly associated with actual transactions. While previous studies have inferred a relationship between ratings and sales through sales rank (Chevalier and Mayzlin, 2003), we show such a direct link through actual transaction data. In addition, our paper is the first to begin to examine retailer involvement in filtering online ratings. Our initial results suggest that consumers use different metrics of online ratings to assess the information depending upon what review information is presented to them. In the situation where the retailer filtered out negative reviews, valence score (average ratings) of the product and of the product category were significantly associated with purchase amount. In addition, consumers seemed to use extreme ratings as a gauge of degree of bias, rather than as an approximation of product quality. However, in the situation where the firm did not filter out negative reviews, consumers appear to use the extreme positive and negative ratings in making their purchase decision. We follow these empirical findings with an analytical assessment of optimal firm filtering strategies. Interestingly, our analytical model finds that retailers benefit when they can selectively filter reviews in favor of more profitable products, suggesting that the bias in the reviews will be less visible to the consumer when it is not uniformly applied across all products. However, when the retailer chooses to either filter or not across the entire website, then it is sometimes optimal not to filter at all. In the next steps of our analysis, we plan to analytically model what we observe in our empirical results: that buyers are adjusting their perception of reviews, and therefore that their demand functions can change based on the retailer's filtering strategy.

REFERENCES

1. Dellarocas, Chris, Awad, Neveen and Xiaoquan (Michael) Zhang. (2004) "Exploring the Value of Online Reviews to Organizations: Implications for Revenue Forecasting and Planning," Proceedings of the International Conference on Information Systems, December, Washington, DC.
2. Chevalier, Judith A., and, Mayzlin, Dina (2003). The Effect of Word of Mouth on Sales: Online Book Reviews. *Yale SOM Working Paper No's. ES-28 & MK-15*.
3. Godes, David, and, Mayzlin, Dina (2002). Using Online Conversations to Study Word of Mouth Communication. *Yale SOM Working Paper No. MK-13; Harvard NOM Working Paper No. 02-32; HBS Marketing Research Paper No. 02-01*.
4. Harmon, Amy (2004). Amazon Glitch Unmasks War of Reviewers. *The New York Times*. New York, February 14.
5. Senecal, S. and Nantel, J. (2003) The Influence of Online Product Recommendations on Consumers' Online Choices. Working Paper.
6. White, E. (1999). Chatting a Singer Up the Pop Charts. *The Wall Street Journal* October 5.