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 in general, 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.2.8 [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

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 aspects 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 second major research contribution, our techniques examine the actual text of the review to identify
which review is expected to have the most impact on sales, and therefore identify the most important reviews from a manufacturer’s point of view.

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 our 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, 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 [19] 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 data set with 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, we follow the same paradigm. In particular, objective information is considered the information that also appears in the product description, and subjective is everything else.

Using this definition, we then generated 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.

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\(^2\).

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 \(P_{\text{subjective}}(s)\) for each sentence \(s\). Hence, for each review, we have a “subjectivity” score for each of the sentences.

\(^1\)We should note, though, that the numeric rating does not capture all the polarity information that appears in the review [1].
\(^2\)http://www.alias-i.com/lingpipe/
Based on the classification scores for the sentences in each review, we derived the average probability \( \text{AvgProb}(r) \) of the review \( r \) being subjective defined as:

\[
\text{AvgProb}(r) = \frac{1}{n} \sum_{i=1}^{n} Pr_{\text{subj}}(s_i)
\]

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 \( \text{DevProb}(r) \) of the subjectivity scores for each review, defined as:

\[
\text{DevProb}(r) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left( Pr_{\text{subj}}(s_i) - \text{AvgProb}(r) \right)^2}
\]

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 \( \text{Read} \) 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 variables as the ratio of the length of the review in characters to the number of sentences.

\[\text{Obs.} \quad \text{Mean} \quad \text{Std. Dev.} \quad \text{Min} \quad \text{Max}\]

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs.</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>AvgProb</td>
<td>18720</td>
<td>.58396</td>
<td>.04195</td>
<td>.37</td>
<td>.8297</td>
</tr>
<tr>
<td>DevProb</td>
<td>18720</td>
<td>.04756</td>
<td>.02378</td>
<td>0</td>
<td>.1807</td>
</tr>
<tr>
<td>Sales Rank</td>
<td>18628</td>
<td>7667.42</td>
<td>51039.42</td>
<td>0</td>
<td>2090308</td>
</tr>
<tr>
<td>Rating</td>
<td>18720</td>
<td>3.8563</td>
<td>1.4141</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Helpful Votes</td>
<td>18720</td>
<td>6.3432</td>
<td>13.873</td>
<td>0</td>
<td>706</td>
</tr>
<tr>
<td>Total Votes</td>
<td>18720</td>
<td>9.6248</td>
<td>16.359</td>
<td>0</td>
<td>847</td>
</tr>
<tr>
<td>Reviews</td>
<td>18616</td>
<td>138.421</td>
<td>202.24</td>
<td>0</td>
<td>1339</td>
</tr>
<tr>
<td>Amazon Price</td>
<td>15108</td>
<td>76.6312</td>
<td>162.73</td>
<td>0</td>
<td>7999.99</td>
</tr>
<tr>
<td>Used Price</td>
<td>16318</td>
<td>116.033</td>
<td>181.58</td>
<td>0</td>
<td>7999</td>
</tr>
<tr>
<td>Num. of Used Goods</td>
<td>12057</td>
<td>39.8082</td>
<td>38.91</td>
<td>0</td>
<td>241</td>
</tr>
<tr>
<td>Sentences</td>
<td>18720</td>
<td>10.3533</td>
<td>10.42</td>
<td>1</td>
<td>160</td>
</tr>
<tr>
<td>Log(Length)</td>
<td>18720</td>
<td>4.5512</td>
<td>0.527</td>
<td>2.681</td>
<td>8.1373</td>
</tr>
<tr>
<td>Log(Elapsed Date)</td>
<td>17000</td>
<td>5.1225</td>
<td>1.095</td>
<td>0</td>
<td>7.6338</td>
</tr>
<tr>
<td>Moderate</td>
<td>18721</td>
<td>.09337</td>
<td>.2909</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1: Descriptive statistics based on all product categories.

where \( \text{AvgProb} \), and \( \text{DevProb} \) are the review-level variables described in Section 3 and \( \mu_k \) 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 date of the product (\( \text{Elapsed Date} \)), the average numeric rating of the product (\( \text{Rating} \)), the number of reviews posted for that product (\( \text{Number of Reviews} \)), and the readability of the review (\( \text{Read} \)). 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:

\[
\delta((\text{SalesRank}_t) - (\text{SalesRank}_{t-1}))
\]

We estimated three different variations of the unit of time: at the daily level, weekly level and bi-weekly level. The results were directionally similar to the ones presented above, and are omitted in this version of the paper for brevity.
only objective information. Compared to reviews that tend to include only subjective or subjective sentences have a positive effect on product sales, that reviews that have a mixture of objective, and highly in sales rank, i.e., an increase in product sales. This means suggests that an increase in deviation leads to a decrease with sales rank in both categories although it is statisti-

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 negatively) are the reviews that should be presented first to the customers. (This is similar to the “spot-lighted” in review forums, like the one of Amazon. We present related evidence next. Effect on Usefulness: 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 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.)

4.2 Effect of Subjectivity on Helpfulness

We use a well-known linear specification for our helpfulness estimation [8]:

$$\ln(HELPFUL)_{kr} = \alpha + \beta_1 \cdot (AvgProb)_{kr} + \beta_2 \cdot (DevProb)_{kr} + \beta_3 \cdot (MODERATE)_{kr} + \beta_4 \cdot \ln(Read)_{kr} + \beta_5 \cdot \ln(ElapsedDate)_{kr} + \mu_k + \epsilon_{kr}$$

where, $k$ and $r$ index product and review. The unit of observation in our analysis is a product-review and $\mu_k$ 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. We also include the MODERATE variable$^3$ to control for the fact that 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].

4.3 Analysis

Effect on Sales: 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. Our conjecture is that customers prefer to read reviews that describe the individual experiences of other consumers and buy products with significant such (subjective) information available.

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.

$^3$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.

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.

Effect on Usefulness: 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 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.)

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

$$\begin{array}{cccc}
\text{Independent Variable} & \text{Audio-Video} & \text{Digital Camera} \\
\hline
\text{AvgProb} & -1.47^{***} (0.72) & 1.27 (1.24) \\
\text{DevProb} & -0.69 (1.06) & -2.91^{***} (1.1) \\
\text{Log (Amazon Price)} & 1.59^{***} (0.3) & 6.2^{***} (0.61) \\
\text{Log(Elapsed Date)} & 0.12 (0.07) & 0.28^{**} (0.13) \\
\text{Average Rating} & -0.01 (0.02) & -0.01 (0.03) \\
\text{Log (Reviews)} & 0.06 (0.15) & 1.08^{***} (0.2) \\
\text{Log(Read)} & 0.06 (0.037) & -0.15^{*} (0.08) \\
K^2 & 0.18 & 0.37 \\
\end{array}$$

5. VALIDATION OF USEFULNESS PREDICTION USING 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:

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

### Table 3: The dependent variable is $\ln(\text{Helpful})$. Standard errors are listed in parenthesis; ***, ** and * denote significance at 1%, 5% and 10%, respectively.

<table>
<thead>
<tr>
<th>Ind. Variable</th>
<th>Audio-Video</th>
<th>Digital Camera</th>
</tr>
</thead>
<tbody>
<tr>
<td>AvgProb</td>
<td>-0.52*** (.28)</td>
<td>-1.9*** (.37)</td>
</tr>
<tr>
<td>DevProb</td>
<td>5.29*** (.42)</td>
<td>4.74*** (.6)</td>
</tr>
<tr>
<td>Log(Elapsed Date)</td>
<td>0.044* (.02)</td>
<td>0.03 (.03)</td>
</tr>
<tr>
<td>MODERATE</td>
<td>-0.13*** (.03)</td>
<td>-0.06 (.04)</td>
</tr>
<tr>
<td>Log(Read)</td>
<td>0.21*** (.019)</td>
<td>0.21*** (.026)</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.07</td>
<td>0.1</td>
</tr>
</tbody>
</table>

### Table 4: F-Measure for the task of predicting whether a review is useful or not. We define a review as useful when useful votes/total votes $\geq 0.6$.

<table>
<thead>
<tr>
<th>Product Category</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio-Video</td>
<td>0.85</td>
</tr>
<tr>
<td>Digital Camera</td>
<td>0.85</td>
</tr>
<tr>
<td>Overall</td>
<td>0.85</td>
</tr>
</tbody>
</table>

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.

In general, our results indicate that we can achieve good empirical performance for the task of usefulness prediction. This means that we can derive from the text both the informativeness 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. Similarly, based on the results from Section 4, we can predict whether a review is expected to have significant effect in the sales. In this case, the manufacturer 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 sentiment analysis and behavioral economics, where sentiment has been shown to significantly affect consumer behavior such as stock purchases and sales. However, sentiment analysis techniques are not yet able to predict the impact of individual sentiment on product sales and the extent to which these sentiments drive decision-making. 

For the effect on sales, we need to conduct further investigations and potentially examine the interactions of the subjective content with the numeric rating of the review. In terms of quality and usefulness, we observe that for feature-based products, such as electronics, users prefer reviews that contain 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 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.

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 when sentiment in text of a review affects product sales 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 subjective content with the numeric rating of the review. 

In terms of quality and usefulness, we observe that for feature-based products, such as electronics, users prefer reviews that contain 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 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


