Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics

Anindya Ghose, Panagiotis G. Ipeirotis, *Member, IEEE,*

Abstract—With the rapid growth of the Internet, the ability of users to create and publish content has created active electronic communities that provide a wealth of product information. However, the high volume of reviews that are typically published for a single product makes harder for individuals as well as manufacturers to locate the best reviews and understand the true underlying quality of a product. In this paper, we re-examine the impact of reviews on economic outcomes like product sales and see how different factors affect social outcomes such as their perceived usefulness. Our approach explores multiple aspects of review text, such as *subjectivity* levels, various measures of *readability* and extent of *spelling errors* to identify important text-based features. In addition, we also examine multiple reviewer-level features such as average usefulness of past reviews and the self-disclosed identity measures of reviewers that are displayed next to a review. Our econometric analysis reveals that the extent of subjectivity, informativeness, readability, and linguistic correctness in reviews matters in influencing sales and perceived usefulness. Reviews that have a mixture of objective, and highly subjective sentences are negatively associated with product sales, compared to reviews that tend to include only subjective or only objective information. However, such reviews are rated more informative (or helpful) by other users. By using Random Forest based classifiers, we show that we can accurately predict the impact of reviews on sales and their perceived usefulness. We examine the relative importance of the three broad feature categories: 'reviewer-related' features, 'review subjectivity' features, and 'review readability' features, and find that using any of the three feature sets results in a statistically equivalent performance as in the case of using all available features. This paper is the first study that integrates econometric, text mining, and predictive modeling techniques toward a more complete analysis of the information captured by user-generated online reviews in order to estimate their helpfulness and economic impact.

1 INTRODUCTION

With the rapid growth of the Internet, product related word-of-mouth conversations have migrated to online markets, creating active electronic communities that provide a wealth of information. Reviewers contribute time and energy to generate reviews, enabling a social structure that provides benefits both for the users and the firms that host electronic markets. In such a context, "who" says "what" and "how" they say it, matters.

On the flip side, a large number of reviews for a single product may also make it harder for individuals to track the gist of users' discussions and evaluate the true underlying quality of a product. Recent work has shown that the distribution of an overwhelming majority of reviews posted in online markets is bimodal [1]. 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 or to the manufacturer who tries to understand what aspects of its product are important. Instead, the reader has to read the actual reviews to examine which of the positive and which of the negative attributes of a product are of interest.

So far, the best effort for ranking reviews for consumers comes in the form of "peer reviewing" in review forums, where customers give "helpful" votes to other reviews in order to signal their informativeness. 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. Similarly, merchants need to know what aspects of reviews are the most informative from consumers' perspective. Such reviews are likely to be the most helpful for merchants, as they contain valuable information about what aspects of the product are driving the sales up or down.

In this paper, we propose techniques for predicting the helpfulness and importance of a review so that we can have:

- a consumer-oriented mechanism which can potentially rank the reviews according to their *expected helpfulness* (i.e., estimating the *social* impact), and
- a manufacturer-oriented ranking mechanism, which can potentially rank the reviews according to their *expected influence on sales* (i.e., estimating the *economic* impact).

To understand better what are the factors that influence consumers perception of usefulness and what factors affect consumers most, we conduct a two-level study. First, we perform an *explanatory* econometric analysis, trying to identify what aspects of a review (and of a reviewer) are important determinants of its usefulness and impact. Then, at the second level we build a *predictive* model using Random Forests that offer significant predictive power and allow us to predict with high accuracy how peer consumers are going to rate a review and how sales will be affected by the posted review.

Our algorithms are based on the idea that the writing style of the review plays an important role in determining the perceived helpfulness by other fellow customers and the degree of influencing purchase decisions. In our work, we perform multiple levels of automatic text analysis to identify characteristics of the review that are important. We

- *Anindya Ghose and Panagiotis G. Ipeirotis are with the Department of Information, Operations, and Management Sciences, Leonarn N. Stern School of Business, New York University, New York, NY, 10012.*
E-mail: {aghose,panos}@stern.nyu.edu.

Manuscript received August 31, 2008; revised on June 30, 2009; revised on Dec 31, 2009.

Digital Object Indentifier 10.1109/TKDE.2010.188

perform our analysis at the *lexical*, *grammatical*, *semantic*, and at the *stylistic* levels to identify text features that have high predictive power in identifying the perceived helpfulness and the economic impact of a review. Furthermore, we examine whether the past history and characteristics of a reviewer can be a useful predictor for the usefulness and impact of a review. We present an extensive experimental analysis using a real data set of 411 products, monitored over a 15-month period on Amazon.com. Our analysis indicates that we can predict accurately the helpfulness and influence of product reviews.

The rest of the paper is structured as follows. First, Section 2 discusses related work and provides the theoretical framework for generating the variables for our analysis. Then, in Section 3, we describe our data set and discuss how we extract the variables that we use to predict the usefulness and impact of a review. In Section 4, we present our explanatory econometric analysis for estimating the influence of the different variables and in Section 5 we describe the experimental results of our predictive modeling that uses Random Forest classifiers. Finally, Section 6 provides some additional discussion and concludes the paper.

2 Theoretical Framework and Related Literature

From a business perspective, consumer product reviews are most influential if they affect product sales and the online behavior of users of the word-of-mouth forum.

2.1 Sales Impact

The first relevant stream of literature assesses the effect of online product reviews on sales. Research in this direction has generally assumed that the primary reason that reviews influence sales is because they provide information about the product or the vendor to potential consumers.

Prior research has demonstrated an association between numeric ratings of reviews (review valence) and subsequent sales of the book on that site [2], [3], [4] or between review volume and sales [5], [6], [7]. Indeed, to the extent that better products receive more positive reviews, there should be a positive relationship between review valence and sales. Research also demonstrated that reviews and sales may be positively related even when underlying product quality is controlled [3], [5].

However, prior work has not looked at how the textual characteristics of a review affect sales. Our hypothesis is that the text of product reviews affects sales even after taking into consideration the numerical information such as review valence and volume. Intuitively, reviews of reasonable length, that are easy to read, and lack spelling and grammar errors should be, all else being equal, more helpful and influential compared to other reviews that are difficult to read and have errors. Reviewers also write “subjective opinions” that portray reviewers’ emotions about product features or more “objective statements” that portray factual data about product features, or a mix of both.

Keeping these in mind, we formulate three potential constructs for text-based features that are likely to have an impact: (i) the average level of subjectivity and the range and mix of subjective and objective comments, (ii) the extent to which the content is easy to read, and (iii) the proportion of spelling errors in the review. In particular, we test the following hypotheses:

Hypothesis 1a*: All else equal, a change in the subjectivity level and mixture of objective and subjective statements in reviews will be associated with a change in sales.*

Hypothesis 1b*: All else equal, a change in the readability score of reviews will be associated with a change in sales.*

Hypothesis 1c*: All else equal, a decrease in the proportion of spelling errors in reviews will be positively related to sales.*

2.2 Helpfulness Votes and Peer Recognition

A second stream of related research on word-of-mouth suggests that perceived attributes of the reviewer may shape consumer response to reviews [5]. In the social psychology literature, message source characteristics have been found to influence judgment and behavior [8], [9], [10], [11], and it has been often suggested that source characteristics might shape product attitudes and purchase propensity. Indeed, Forman et al. [5] draw on the information processing literature to suggest that product sales will be affected by reviewer disclosure of identity-related information. Prior research on computer mediated communication (CMC) suggests that online community members communicate information about product evaluations with an intent to influence others’ purchase decisions as well as provide social information about contributing members themselves [12], [13]. Research concerning the motivations of content creators in online contexts highlights the role of identity motives in defining why users provide social information about themselves (e.g., [14], [15], [16], [17]).

Increasingly, we have seen that both identity-descriptive information about reviewers and product information is prominently displayed on the web sites of online retailers. Prior research on self-verification in online contexts has pointed out the use of persistent labeling, defined as using a single, consistent way of identifying oneself such as ’real name’ in the Amazon context, and self-presentation, defined as ways of presenting oneself online that may help others to identify one, such as posting geographic location or a personal profile in the Amazon context [17] as important phenomena in the online world. Indeed, information about product reviewers is often highly salient. Visitors to the site can see more professional aspects of reviewers such as their badges (e.g., “top-50 reviewer,” “top-100 reviewer” badges) and ranks (“reviewer rank”) as well as personal information about reviewers ranging from their real name to where they live, their nick names, hobbies, professional interests, pictures and other posted links. In addition, users have the opportunity to examine more “professional” aspects of a reviewer such as the proportion of helpful votes given by other users not only for a given review but across all the reviews of all other products posted by a reviewer. Further, interested users can also read the actual content of all reviews generated by a reviewer across all products.

With regard to the benefits reviewers derive, work on online user-generated content has primarily focused on the consequences of peer recognition rather than on its antecedents [18], [19]. Its only recently that [5] evaluated the influence of reviewers’ disclosure of information about themselves on the extent of peer recognition of reviewers and their interactions with the review valence by drawing on the social psychology literature. We hypothesize that after controlling for features examined in prior work such as reviewer disclosure of identity information and the valence of reviews, the actual text of the review matters in determining the extent to which users find the review useful. In particular, we focus on four constructs, namely subjectiveness, informativeness, readability, and proportion of spelling errors. Our paper thus contributes to the existing stream of work by examining text-based antecedents of peer recognition in online word-of-mouth forums. In particular, we test the following hypotheses:

Hypothesis 2a*: All else equal, a change in the subjectivity level and mixture of objective and subjective statements in a review will be associated with a change in the perceived helpfulness of that review.*

Hypothesis 2b*: All else equal, a change in the readability of a review will be associated with a change the perceived helpfulness of that review.*

Hypothesis 2c*: All else equal, a decrease in the proportion of spelling errors in a review will be positively related to perceived helpfulness of that review.*

Hypothesis 2d*: All else equal, an increase in the average helpfulness of a reviewer's historical reviews will be positively related to perceived helpfulness of a review posted by that reviewer.*

This paper builds on our previous work [20], [21], [22]. In [20], [21] we examined just the effect of subjectivity, while in the current work, we expanding our data to include more product categories and examine a significantly increased number of features, such as different readability metrics, information about the reviewer history, different features of reviewer disclosure and so on. The present paper is unique in looking at how various additional features of the review text affects product sales and the perceived helpfulness of these reviews.

In parallel with our work, researchers in the natural language processing field have examined the task of predicting review helpfulness [23], [24], [25], [26], [27], using reviews from Amazon.com or movie reviews as training and test data. Our work uses a superset of the features used in the past for helpfulness prediction (e.g., reviewer history and disclosure, deviation of subjectivity in the review, and so on). Also, none of these studies attempts to predict the influence of reviews on product sales. A differentiating factor of our approach is the two-pronged approach building on methodologies from economics and from data mining, building both *explanatory and predictive models* to understand better the impact of different factors. Interestingly, all prior research use support vector machines (in a binary classification and in regression mode), which we observed to perform worse than Random Forests (as we discuss in Section 5). Predicting the helpfulness of a review is also related to the task of evaluating the quality of web posts or the quality of answers to posted questions [28], [29], [30], [31], although there are more cues (e.g., clickstream data) that can be used to estimate the perceived quality of a posting. Recently, Hao et al. [32] also presented techniques for predicting whether a review will receive any votes about its helpfulness or whether it will stay unrated. Tsur and Rappoport [33] presented an unsupervised algorithm for estimating ranking the reviews according to their expected helpfulness.

3 Data Set and Variables

A major goal of this paper is to explore how the user-generated textual content of a review and the self-reported characteristics of the reviewer who generated the review can influence economic transactions (such as product sales) and online community and social behavior (such as peer recognition in the form of helpful votes). To examine this, we collected data about the economic transactions on Amazon.com and analyzed the associated review system. In this section, we describe the data that we collected from Amazon; furthermore, we discuss how we computed the variables to perform our analysis, based on the discussion of Section 2.

3.1 Product and Sales Data

To conduct our study, we created a panel data set of products belonging to three product categories:

1) Audio and video players (144 products),
2) Digital cameras (109 products), and
3) DVDs (158 products).

We picked the products by selecting all the items that appeared in the "Top-100" list of Amazon over a period of 3 months from January 2005 to March 2005. We decided to use popular products, in order to have products in our study with a significant number of reviews. Then, using Amazon web services, from March 2005 until May 2006 we collected the information for these products described below.

We collected various product-specific characteristics over time. Specifically, we collected the manufacturer suggested *list price* of the product, its Amazon *retail price*, its Amazon *sales rank* (which serves as a proxy for units of demand [34], as we will describe later).

Together with sales and price data, we also collected other data that may influence the purchasing behavior of consumers. For example, we collected the date the product was released into the market, to compute the *elapsed time* from the date of product release, since products released long time ago tend to see a decrease in sales over time. We also collected the *number of reviews* and the *average review rating* of the product over time.

3.2 Individual Review Data

Beyond the product-specific data, we also collected all reviews of a product since the product was released into the market. For each review, we retrieve the actual textual content of the review and the *review rating* of the product given by the reviewer. The rating that a reviewer allocates to the reviewed product is denoted by a number of stars on a scale of 1 to 5. From the textual content, we generated a set of variables at the lexical, grammatical, and at the stylistic level. We describe these variables in detail in Section 3.4, when we describe the textual analysis that we conducted.

Review Helpfulness: Amazon has a voting system whereby community members provide helpful votes to rate the reviews of other community members. Previous peer ratings appear

Type	**Variable**	**Explanation**
Product and Sales Data	Retail Price	The retail price at Amazon.com
	Sales Rank	The sales rank within the product category
	Average Rating	Average rating of the posted reviews
	Number of Reviews	Number of reviews posted for the product
	Elapsed Date	Number of days since the release of the product
Individual Review	Moderate Review	Does the Review have a moderate rating (3 star rating) or not
	Helpful Votes	The number of helpful votes for the review
	Total Votes	The total number of votes for the review
	Helpfulness	$\frac{HelpfulVotes}{TotalVotes}$
Reviewer Characteristics	Reviewer Rank	The reviewer rank according to Amazon
	Top-10 Reviewer	Is the reviewer a Top-10 reviewer?
	Top-50 Reviewer	Is the reviewer a Top-50 reviewer?
	Top-100 Reviewer	Is the reviewer a Top-100 reviewer?
	Top-500 Reviewer	Is the reviewer a Top-500 reviewer?
	Real Name	Has the reviewer disclosed his/her real name?
	Nickname	Does the reviewer have a nickname listed in the profile?
	Hobbies	Does the reviewer have an ”about me” section in the profile?
	Birthday	Does the reviewer list his/her birthday?
	Location	Does the reviewer disclose its location?
	Web Page	Does the reviewer have a home page listed?
	Interests	Does the reviewer list his/her interests?
	Snippet	Does the reviewer has a description in the reviewer profile?
	Any Disclosure	Does the reviewer list *any of the above* in the reviewer profile?
Reviewer History	Number of Past Reviews	Number of reviews posted by the reviewer
	Reviewer History Macro	Average past review helpfulness (macro-averaged)
	Reviewer History Micro	Average past review helpfulness (micro-averaged)
	Past Helpful Votes	Number of helpful votes accumulated in the past from the reviewer
	Past Total Votes	Number of total votes on the reviews posted in the past for the reviewer
Review Readability	Length (Chars)	The length of the review in characters
	Length (Words)	The length of the review in words
	Length (Sentences)	The length of the review in sentences
	Spelling Errors	The number of spelling errors in the review
	ARI	The Automated Readability Index (ARI) for the review
	Gunning Index	The Gunning–Fog index for the review
	Coleman–Liau index	The Coleman–Liau index for the review
	Flesch Reading Ease	The Flesch Reading Ease score for the review
	Flesch–Kincaid Grade Level	The Flesch–Kincaid Grade Level for the review
	SMOG	The Simple Measure of Gobbledygook score for the review
Review Subjectivity	AvgProb	The average probability of a sentence in the review being subjective
	DevProb	The standard deviation of the subjectivity probability

TABLE 1

The variables collected for our study. The panel data set contains data collected over a period of 15 months; we collected the variables daily and we capture the variability over time for the variables that change over time (e.g., sales rank, price, reviewer characteristics and so on).

immediately above the posted review, in the form, “[*number of helpful votes*] out of [*number of members who voted*] found the following review helpful.” These *helpful* and *total* votes enable us to compute the fraction of votes that evaluated the review as helpful. To have as much accurate representation of the percentage of customers that found the review helpful, we collected the votes in December 2007, ensuring that there is a significant time period after the time the review was posted and that there is a significant number of peer rating votes accumulated for the review.

3.3 Reviewer Characteristics

Reviewer Disclosure: While review valence is likely to influence consumers, there is reason to believe that social information about reviewers themselves (rather than the product or vendor) is likely to be an important predictor of consumers’ buying decisions [5]. On many sites, social information about the reviewer is at least as prominent as product information. For example, on sites such as Amazon, information about product reviewers is graphically depicted, highly salient, and sometimes more detailed and voluminous than information on the products they review: the “Top-1000” reviewers have special tags displayed next to their names, the reviewers that disclose their real name[1] are also highlighted and so on. Given the extent and salience of available social information regarding product reviewers, it seems important to control for the impact of such information on online product sales and review helpfulness. Amazon has a procedure by which reviewers can disclose personal information about themselves. There are several types of information that users can disclose: we focus our analysis on the categories most commonly indicated by users: whether the user disclosed their *real name*, their *location*, *nickname*, and *hobbies*. With real name, we refer to a registration procedure that Amazon provides for users to indicate their actual name by providing verification with a credit card, as mentioned above. Reviewers

[1] Amazon compares the name of the reviewer with the name listed in the credit card on file before assigning the “Real Name” tag.

may also post additional information in their profiles such as geographical location, disclose additional information (e.g., "Hobbies") or use a nickname (e.g., "Gadget King"). We use these data to control for the impact of self-descriptive identity claims. We encode this information as binary variables. We also constructed an additional dummy variable, labeled "*any disclosure*"; this variable captures each instance where the reviewer has engaged in any one of the four kinds of self-disclosure. We also collected the *reviewer rank* of the reviewer as published on Amazon.

Reviewer History: Since one of our goal is to predict the future usefulness of a review, we wanted to examine whether the past history of a reviewer can be used to predict the usefulness of the future reviews written by the same reviewer. For this, we collected the past reviews for each reviewer, and collected the helpful and total votes for each of the past reviews. Using this information, we constructed *for each reviewer and for each point in time* the past performance of a reviewer. Specifically, we created two variables, by *micro-averaging* and *macro-averaging* the past votes on the reviews. The variable *reviewer history macro*, is the ratio of all past helpful votes divided by the total number of votes. Similarly, we also created the variable *reviewer history micro*, in which we first computed the average helpfulness for each of the past reviews and then computed the average across all past reviews. The difference with the macro and micro versions is that the micro version gives equal weight to the helpfulness of all past reviews, while the macro version weights more heavily the importance of reviews that received a large number of votes.

3.4 Textual Analysis of Reviews

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 [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46]. 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.

Readability Analysis: We are interested to examine what types of reviews affect most sales and what types of reviews are most helpful to the users. For example, everything else being equal, a review that is easy to read will be more helpful than another that has spelling mistakes and is difficult to read.

As a first, low-level variable, we measured the number of spelling mistakes within each review, and we normalized the number by dividing with the length of the review (in characters).[3] To measure the spelling errors, we used an off-the-shelf spell checker, ignoring capitalized words and words with numbers in them. We also ignored the top-100 most frequent non-English words that appear in the reviews: most of them were brand names or terminology words that do not appear in the spellcheckers list. Furthermore, to measure the cognitive effort that a user needs in order to read a review, we measured the length of a review in sentences, words, and characters.

Beyond these basic features, we also used the extensive results from research on *readability*. Past research has shown that easy-reading text improves comprehension, retention, and reading speed, and that the average reading level of the US adult population is at the eighth grade level [47]. Therefore, a review that can be read easily by a large number of users is also expected to be rated by more users. Today there are numerous metrics for measuring the readability of a text, and while none of them is perfect, the computed measures correlate well with the actual difficulty of reading a text. To avoid idiosyncratic errors peculiar to a specific readability metric, we computed a set of metrics for each review. Specifically, we computed the following:

- Automated Readability Index
- Coleman–Liau Index
- Flesch Reading Ease
- Flesch–Kincaid Grade Level
- Gunning fog index
- SMOG

(See [48] for detailed description on how to compute each of these metrics.) Based on research in readability, these metrics are useful metrics for measuring how easy is for a user to read a review.

Subjectivity Analysis: Beyond the lower level spelling and readability analysis, we also expect that there are stylistic choices that affect the perceived helpfulness of a review. We observed empirically that there are two types of listed information, 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 style of the reviews on helpfulness and product sales, we rely on existing literature of subjectivity estimation from computational linguistics [41]. Specifically, Pang and Lee [41] 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 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

[2] We should note, though, that the numeric rating does not capture all the polarity information that appears in the review [19].

[3] To take the logarithm of the normalized variable for errorless reviews, we added one to the number of spelling errors before normalizing.

Variable	Obs.	Mean	Std. Dev.	Min	Max
Retail Price	5699	151.33	130.57	0	3299.99
Sales Rank	7352	7667.42	51039.42	0	2090308
Log (Elapsed Date)	7352	5.12	1.09	0	7.63
Average Rating	7352	3.86	1.41	1	5
Number of Reviews	7352	195.07	138.76	0	522
Moderate Review	7352	0.093	0.29	0	1
Any Disclosure	7352	0.52	0.49	0	1
Helpful Votes	7352	5.51	11.7	0	744
Total Votes	7352	8.38	14.05	0	893
Log(Spelling Errors)	7352	-3.85	0.74	-6.67	-1.34
Readability (Gunning)	7352	12.46	13.31	1.36	277.95
AvgProb	7352	0.58	0.05	0.37	0.83
DevProb	7352	0.047	0.024	0	0.18
Rev. History Macro	3076	0.69	0.23	0	1

TABLE 2
Descriptive statistics of audio and video players for Econometric Analysis

Variable	Obs.	Mean	Std. Dev.	Min	Max
Retail Price	1690	159.94	351.62	0	7999.99
Sales Rank	2730	395.27	1418.04	0	38353
Log(Elapsed Date)	2730	5.46	0.55	1.38	7.02
Average Rating	2730	4.06	1.31	1	5
Number of Reviews	2730	84.69	492.62	0	3795
Moderate Review	2730	0.36	0.48	0	1
Any Disclosure	2730	0.58	0.49	0	1
Helpful Votes	2730	17.68	58.58	0	1669
Total Votes	2730	21.67	59.68	0	1689
Log(Spelling Errors)	2730	-4.23	0.86	-7.03	-0.84
Readability (Gunning)	2730	13.07	8.25	1.2	117.44
AvgProb	2730	0.55	0.02	0.46	0.67
DevProb	2730	0.03	0.01	0.001	0.17
Rev. History Macro	1085	0.78	0.22	0	1

TABLE 3
Descriptive statistics of digital cameras for Econometric Analysis

Variable	Obs.	Mean	Std. Dev.	Min	Max
Retail Price	2005	21.30	7.81	0	71.96
Sales Rank	2018	3824.28	6168.97	6	49868
Log (Elapsed Date)	2018	4.38	1.29	0	7.55
Average Rating	2018	4.05	0.62	1	5
Number of Reviews	2018	84.54	88.36	0	431
Moderate Review	2018	0.45	.49	0	1
Any Disclosure	2018	0.74	0.43	0	1
Helpful Votes	2018	6.35	13.626	0	246
Total Votes	2018	10.498	19.31	0	320
Log(Spelling Errors)	2018	-4.34	0.91	-7.44	-1.003
Readability (Gunning)	2018	13.09	6.47	0.8	136.83
AvgProb	2018	0.52	0.012	0.46	0.56
DevProb	2018	0.016	.01	0.00003	0.07
Rev. History Macro	1450	0.68	0.21	0	1

TABLE 4
Descriptive statistics of DVD for Econometric Analysis

toolkit[4]. The accuracy of the classifiers according to the Area under the ROC curve, measured using 10-fold cross validation was: 0.85 for audio and video players, 0.87 for digital cameras, and 0.82 for DVDs.

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 the mean value of the $Pr_{subj}(s_i)$ values for the sentences $s_1, \ldots, s_n$ in the 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 $Pr_{subj}(s_i)$ for the sentences in each review.[5]

The summary statistics of the data for audio-video players, digital cameras and DVDs are given in Table 2, Table 3 and Table 4, respectively.

4 Explanatory Econometric Analysis

So far, we have explain the different types of data that we collected, that have the potential, according the various hypotheses, to affect the impact and usefulness of the reviews. In this section, we present the results of our *explanatory* econometric analysis, which will examine the importance of each factor. Through our analysis, we aim to provide a better understanding of how customers are affected by the reviews. (In the next section, we will describe our *predictive* model, based on machine learning techniques.) In Section 4.1 we analyze the effect of different review and reviewer characteristics on product sales. Our results show what factors are important for a merchant to observe. Then, in Section 4.2 we presents our analysis on how different factors affect the helpfulness of a review.

4.1 Effect on Product Sales

We first estimate the relationship between sales and stylistic elements of a review. Prior research in economics and in marketing (for instance, [49]) has associated sales ranks with demand levels for products such as software and electronics. The association is based on the experimentally observed fact that the distribution of demand in terms of sales rank has a Pareto distribution (i.e., a power law). Based on this observation, it is possible to convert sales ranks into demand levels using the following Pareto relationship:

$$\ln(D) = a + b \cdot \ln(S) \tag{1}$$

where D is the unobserved product demand, S is its observed sales rank, and $a > 0$, $b < 0$ are industry-specific parameters. Therefore, we can use the log of product sales rank on Amazon.com as a proxy of the log of product demand.

Previous work has examined how price, number of reviews, and review valence influence product sales on Amazon and Barnes and Noble [3]. Recent work by Forman et al. [5] also describes how reviewer disclosure of identity descriptive

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

[5] To examine the extent to which people cross-reference reviews such as " I agree with Tom", we did an additional study. We posted 2000 product reviews on Amazon Mechanical Turk, asking workers there to examine the reviews and indicate whether the reviewer refers to some other review. We asked 5 workers on Mechanical Turk to annotate each review. If at least one worker indicated that the review refers to some other review or webpage, then we classified a review as "cross-referencing". The extent of cross-referencing was very small. Out of the 2000 reviews only 38 had at least one "cross-referencing" vote (1.9%), and only 2 reviews were judged as "cross-referencing" by all 5 workers (0.1%). This corresponds to a relatively limited source of errors and does not affect significantly the accuracy of the subjectivity classifier.

information (e.g., *Real Name* or *Location*) affects product sales. Hence, to be consistent with prior work, we control for all these factors but focus mainly on the textual aspects of the review to see how they affect sales.

4.1.1 Model Specification

In order to test our hypotheses 1a to 1c, we adopt a model similar to that used in [3] and [5], while incorporating measures for the quality and the content of the reviews. Chevalier and Mayzlin [3] and Forman, Ghose and Wiesenfeld [5] 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. The dependent variable is $\ln(SalesRank)_{kt}$, the log of sales rank of product k in time t, which is a linear transformation of the log of product demand, as discussed earlier. The unit of observation in our analysis is a product-date: since we only know the date that a review is posted (and not its time) and we observe changes in sales rank on a daily basis, we need to "collapse" multiple reviews posted on the same data in a single observation. Since we have a linear model, we use an additive approach to combine reviews published for the same product on the same date. To study the impact of reviews and the quality of reviews on sales, we estimate the following model:

$$\begin{aligned}\log(SalesRank)_{kt} = \alpha + &\beta_1 \cdot \log(AmazonPrice_{kt}) + \\ &\beta_2 \cdot AvgProb_{k(t-1)} + \\ &\beta_3 \cdot DevProb_{k(t-1)} + \\ &\beta_4 \cdot AverageReviewRating_{k(t-1)} + \\ &\beta_5 \cdot \log(NumberofReviews_{k(t-1)}) + \\ &\beta_6 \cdot (Readability_{k(t-1)}) + \\ &\beta_7 \cdot \log(SpellingErrors_{k(t-1)}) + \\ &\beta_8 \cdot (AnyDisclosure_{k(t-1)}) + \\ &\beta_9 \cdot \log(ElapsedDate_{kt}) + \\ &\mu_k + \varepsilon_{kt} \end{aligned} \tag{2}$$

where μ_k is a product *fixed effect* that accounts for unobserved heterogeneity across products and ε_{kt} is the error term. (The other variables are described in Table 1 and in Section 3.) To select the variables that are present in the regression, we follow the work in [3], [5].[6]

Note that as explained above increases in sales rank mean lower sales, so a negative coefficient on a variable implies that an increase in that variable increases sales. The control variables used in our model include the Amazon retail 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 (*Rating*), and the log of the number of reviews posted for that product (*Number of Reviews*). This is consistent with prior work such as Chevalier and Mayzlin [3] and Forman et al. [5] To account for potential non-linearities and to smooth large values, we take the log of the dependent variable and some of the control variables such as Amazon Price, volume of reviews and days elapsed consistent with the literature [5], [34]. For these regressions in which we examine the relationship between review sentiments and product sales, we aggregate data to the weekly level. By aggregating data in this way, we smooth potential day-to-day volatility in sales rank. (As a robustness check, we also ran regressions at the daily and fortnightly level, and find that the qualitative nature of most of our results remain the same.)

We estimate product-level fixed effects to account for differences in average sales rank across products. These fixed effects are algebraically equivalent to including a dummy for every product in our sample, and so this enables us to control for differences in the average quality of products. Thus, any relationship between sales rank and review valence will not reflect differences in average quality across products, but rather will be identified off changes over time in sales rank and review valence within products, diminishing the possibility that our results reflect differences in average unobserved book quality rather than aspects of the reviews themselves [5].

Our primary interest is in examining the association between textual variables in user-generated reviews and sales. To maintain consistency with prior work, we also examine the association between average review valence and sales. However, prior work has shown that review valence may be correlated with product-level unobservables that may be correlated with sales. In our setting, though we control for differences in the average quality of products through our fixed effects, it is possible that changes in the popularity of the product over time may be correlated with changes in review valence. Thus, this parameter reflects not only the information content of reviews but also may reflect exogenous shocks that may influence product popularity [5]. Similarly, the variable *Number of Reviews* will also capture changes in product popularity or perceived product quality over time; thus, β_5 may reflect the combined effects of a causal relationship between number of reviews and sales [7] and changes in unobserved book popularity over time.[7]

4.1.2 Empirical Results

The sign on the coefficient of *AvgProb* suggests that an increase in the average subjectivity of reviews leads to an increase in sales for products, although the estimate is statistically significant only for audio-video players and digital cameras (see Table 5). It is statistically insignificant for DVDs. 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 only for *search goods* (such as cameras and audio-

[6] To avoid accidental discovery of "important" variables, and model building, we have performed a large number of statistical significance tests. Towards a systematic process of variable selection for our regressions, we used the well known *stepwise regression* method. This is a sequential process for fitting the least squares model, where at each step a single explanatory variable is either added to or removed from the model in the next fit. The most commonly used criterion for the addition or deletion of variables in stepwise regression is based on the $partialF - statistic$ for each of the regressions which allows one to compare any reduced (or empty) model to the full model from which it is reduced.

[7] Note that prior work in this domain has often transformed the dependent variable (sales rank) into quantities using the specification similar to Ghose et al. [34]. That was usually done because those papers were interested in demand estimation and imputing price elasticities. However, in this case we are not interested in estimating demand, and hence we do not need to make the actual transformation. In this regard, our paper is more closely related to [5].

Variable	Audio Video	Digital Camera	DVD
Retail Price	0.095 (0.007)***	0.099 (0.01)***	0.212 (0.051)**
AvgProb	-1.49 (0.77)*	-0.163 (1.5)	-2.66 (1.68)
DevProb	2.31 (1.03)**	-0.689 (2.08)	5.24 (2.59)**
Average Rating	-0.32 (0.078)***	0.15 (0.11)	-0.04 (0.11)
Log(Number of Reviews)	-0.618 (0.042)	-0.298 (0.066)***	-0.33 (0.08)***
Readability	0.002 (0.004)	-0.014 (0.006)**	0.004 (0.005)
Log(Spelling Errors)	-0.012 (0.024)	-0.015 (0.034)	0.041 (0.024)*
Any Disclosure	-0.047 (0.030)	-0.023 (0.056)	-0.036 (0.049)
Log(Elapsed Date)	0.120 (0.070)	0.255 (0.12)**	0.831 (0.032)***
Number of Observations	5699	1690	2005
R-square (with fixed effects)	0.91	0.77	0.87

TABLE 5

These are OLS regressions with product-level fixed effects. The dependent variable is Log (Salesrank). Robust standard errors are listed in parenthesis; ***, ** and * denote significance at 1%, 5% and 10%, respectively. The R-square includes fixed effects in R-square computation.

video players) but not for *experience goods*.[8]

The coefficient of *DevProb* has a positive and statistically significant relationship with sales rank in audio-video players and DVDs, but is statistically insignificant for digital cameras. In general this suggests that a decrease in the deviation of the probability of subjective comments 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 negative effect on product sales, compared to reviews that tend to include only subjective or only objective information.

The coefficient of the *Readability* is negative and statistically significant for digital cameras suggesting that reviews that have higher Readability scores are associated with higher sales. This is likely to happen if such reviews are written in more authoritative and sophisticated language which enhances the credibility and informativeness of such reviews. Our results are robust to the use of other Readability metrics described in Table 1 such as ARI, Coleman–Liau index, Flesch–Reading Ease, Flesch–Kincaid Grade Level and the SMOG index.

The coefficient of *Spelling Errors* is positive and statistically significant for DVDs suggesting that an increase in the proportion of spelling mistakes in the content of the reviews decreases product sales for some products whose quality can be assessed only after purchase. However, for hedonic like products such as audio-video players and digital cameras whose quality can be assessed prior to purchase, the proportion of spelling errors in reviews does not have a statistically significant impact on sales. For all three categories, we find that this result is robust to different specifications of normalizing the number of spelling errors such as normalizing by the number of characters, words or sentences in a given review. In sum, our results provide support for Hypotheses 1a to 1c.

As expected, our control variables suggest that sales decrease as Amazon's price increases. Further, even though the coefficient of *Any Disclosure* is statistically insignificant, the negative sign implies that the prevalence of reviewer disclosure of identity-descriptive information would be associated with higher subsequent sales. This is consistent with prior research in the information processing literature supporting a direct effect for source characteristics on product evaluations and purchase intentions when information is processed heuristically [5]. Our results are robust to the use of other disclosure variables in the above regression. For example, instead of "Any Disclosure", if we were to use disclosures of the two most salient reviewer self-descriptive features (Real Name and Location), results are generally consistent with the existing ones.

We also find that an increase in the volume of reviews is positively associated with sales of DVDs and digital cameras. In contrast, average review valence has a statistically significant effect on sales for only audio-video players. These mixed findings are consistent with prior research which have found a statistically significant effect of review valence but not review volume on sales [3], and with others who have found a statistically significant effect of review volume but not valence on sales [5], [7], [6]. [9]

Finally, we also ran regressions that included interaction terms between ratings and the textual variables like *AvgProb*, *DevProb*, *Readability*, and *Spelling Errors*. For brevity, we cannot include these results in the paper. However, a counter-intuitive theme that emerged is that reviews that rate products *negatively* (ratings $<= 2$) can be associated with *increased product sales* when the review text is *informative and detailed* based on its readability score, number of normalized spelling errors, or the mix of subjective and objective sentences). This is likely to occur when the reviewer clearly outlines the pros and cons of the product, thereby providing sufficient information to the consumer to make a purchase. If the negative attributes of the product do not concern the consumer as much as it did the reviewer, then such informative reviews can lead to increased sales.

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

[8] Search goods are those whose quality can be observed before buying the product (e.g., electronics) while for experience goods, the consumers has to consume/experience the product in order to determine its quality (e.g., books, movies).

[9] Note that we do not have other variables such as "Reviewer Rank" or " Helpfulness" in this regression because of a concern that these variables will lead to biased and inconsistent estimates. Said simply, it is entirely possible that 'Reviewer Rank" or " Helpfulness" is correlated with other unobserved review-level attributes. Such correlations between regressors and error terms will lead to the well known endogeneity bias in OLS regressions [50].

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 on Helpfulness

Next, we want to analyze the impact of review variables on the extent to which community members would rate reviews helpful after controlling for the presence of self-descriptive information. Recent work [5] describes how reviewer disclosure of identity descriptive information and the extent of equivocality of reviews (based on the review valence) affects perceived usefulness of reviews. Hence, to be consistent with prior work, we control for these factors but focus mainly on the textual aspects of the review and the reviewer history to see how they affect the usefulness of reviews.[10]

4.2.1 Model Specification

The dependent variable, $Helpfulness_{kr}$, is operationalized as the ratio of helpful votes to total votes received for a review *r* issued for product *k*. In order to test our hypotheses 2a to 2d, we use a well-known linear specification for our helpfulness estimation [5]:

$$\begin{aligned}\log(Helpfulness)_{kr} = {} & \alpha + \beta_1 \cdot (AvgProb)_{kr} + \\ & \beta_2 \cdot (DevProb)_{kr} + \\ & \beta_3 \cdot (AnyDisclosure)_{kr} + \\ & \beta_4 \cdot (Readability)_{kr} + \\ & \beta_5 \cdot (ReviewerHistoryMacro)_{kr} + \\ & \beta_6 \cdot \log(SpellingErrors)_{kr} + \\ & \beta_7 \cdot (Moderate)_{kr} + \\ & \beta_8 \cdot \log(NumberofReviews)_{kr} + \\ & \mu_k + \varepsilon_{kr} \qquad (3)\end{aligned}$$

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 and ε_{kt} is the error term. (The other variables are described in Table 1 and in Section 3.) We also constructed a dummy variable to differentiate between extreme reviews, which are unequivocal and therefore provide a great deal of information to inform purchase decisions, and moderate reviews which provide less information. Specifically, ratings of 3 were classified as *Moderate* reviews while ratings nearer the endpoints of the scale (1, 2, 4, 5) were classified as unequivocal [5].

The above equation can be estimated using a simple panel data fixed effects model. However, one concern with this strategy is that the posting of personal identity information such as Real Name or location may be correlated with some unobservable reviewer-specific characteristics that may influence review quality [5]. If some explanatory variables are correlated with errors, then ordinary least squares regression gives biased and inconsistent estimates. To control for this potential problem, we use a Two Stage Least Squares (2SLS) regression with instrumental variables [50]. Under the 2SLS approach, in the first stage, each endogenous variable is regressed on all valid instruments, including the full set of exogenous variables in the main regression. Since the instruments are exogenous, these approximations of the endogenous covariates will not be correlated with the error term. So, intuitively they provide a way to analyze the relationship between the dependant variable and the endogenous covariates. In the second stage, each endogenous covariate is replaced with its approximation estimated in the first stage and the regression is estimated as usual. The slope estimator thus obtained is consistent [50].

Specifically, we instrument for in the above equation using lagged values of disclosure, subjectivity and readability. We experimented with different combinations of these instruments and find that the qualitative nature of our results are generally robust. The intuition behind the use of these instrument variables is that they are likely to be correlated with the relevant independent variables but uncorrelated with unobservable characteristics that may influence the dependent variable. For example, the use of a Real Name in prior reviews is likely to be correlated with the use of Real Name in the subsequent reviews but uncorrelated with unobservables that determine perceived helpfulness for a given review. Similarly, the presence of subjectivity in prior reviews is likely to be correlated with the presence of subjectivity in subsequent reviews but unlikely to be correlated with the error term that determines perceived helpfulness for the current review. Hence, these are valid instruments in our 2SLS estimation. This is consistent with prior work [5].

To ensure that our instruments are valid, we conducted the Sargan Test of overidentifying restrictions [50]. The joint null hypothesis is that the instruments are valid instruments, i.e., uncorrelated with the error term, and that the excluded instruments are correctly excluded from the estimated equation. For the 2SLS estimator, the test statistic is typically calculated as N*R-squared from a regression of the IV residuals on the full set of instruments. A rejection casts doubt on the validity of the instruments. Based on the p-values from these tests, we are unable to reject the null hypothesis for each of the three categories, thereby confirming the validity of our instruments.

4.2.2 Empirical Results

With regard to the usefulness of reviews, our analysis reveals that for product categories such as audio and video equipments, digital cameras, and DVDs the extent of subjectivity in a review has a statistically significant effect on the extent to which users perceive the review to be helpful. The coefficient of *AvgProb* is negative suggesting that highly subjective reviews are rated as being less helpful. Although *DevProb* is statistically significant for audio video products only, it always has a positive relationship with helpfulness votes. This results suggests that consumers find reviews that have a wide

[10] We compared our work with the model used in [5] who estimated sales but without incorporating the impact of review text variables (such as *AvdProb, DevProb, Readability,* and *Spelling Errors*) and without incorporating the reviewer-level variables (such as *Reviewer History Macro* and *Reviewer History Micro*). There is a significant improvement in R-squared for each model. Specifically, for the regressions used to estimate the impact on product sales (equation 2), our model increases R-squared by 9% for audio-video products, 14% for digital cameras and 15% for DVDs.

Variable	**Audio Video**	**Digital Camera**	**DVD**
AvgProb	-1.184 (0.34)***	-1.284 (0.29)***	-1.440 (0.840)*
DevProb	0.77 (0.41)*	0.33 (0.3)	1.320 (0.950)
Disclosure	0.210 (0.12)*	0.374 (0.119)***	-0.360 (0.240)
Readability	0.003 (0.001)**	-0.001 (0.001)	0.016 (0.004)***
Reviewer History Macro	0.031 (0.035)	-0.063 (0.038)*	0.230 (0.060)***
Log (Spelling Errors)	-0.037 (0.006)***	0.010 (0.016)	-0.040 (0.010)***
Moderate	-0.01 (0.01)	-0.119 (0.018)***	-0.03 (0.018)***
Log (Number of reviews)	0.001 (0.003)	0.024 (0.008)***	0.01 (0.004)***
Number of Observations	3076	1085	1450
R-square	0.08	0.02	0.03

TABLE 6

These are 2SLS regressions with Instrument Variable. Fixed effects are at the product-level. The dependent variable is $Helpful$. Robust standard errors are listed in parenthesis; ***, ** and * denote significance at 1%, 5% and 10%, respectively. The p-values from the Sargan test of overidentifying restrictions confirm the validity of instruments.

range of subjectivity/objectivity scores across sentences to be more helpful. In other words, reviews that have a mixture of sentences with objective and of sentences with extreme, subjective content are rated highly by users. It is worthwhile to mention that we observed the opposite effect for product sales, indicating that helpful reviews are not necessarily the ones that lead to increases in sales.

The negative and statistically significant sign on the coefficient of the *Moderate* variable for two of the three product categories implies that as the content of the review becomes more moderate or equivocal, the review is considered less helpful by users. This result is consistent with the findings of [5] who analyze a panel of book reviews and find a similar negative relationship between equivocal reviews and perceived helpfulness. Increased disclosure of self-descriptive information *Disclosure* typically leads to more helpful votes as can be seen for audio-video players and digital cameras.

We also find that for audio-video players and DVDs, a higher readability score *Readability* is associated with a higher percentage of helpful votes. As with sales, these results are robust to the use of other Readability metrics described in Table 1 such as ARI, Coleman–Liau index, Flesch–Reading Ease, Flesch–Kincaid Grade Level and the SMOG index. In contrast, an increase in the proportion of spelling errors *Spelling Errors* is associated with a lower percentage of helpful votes for both audio-video players and DVDs. For all three categories, we find that this result is robust to different specifications of normalizing the number of spelling errors such as normalizing by the number of characters, words or sentences in a given review. Finally, the past historical information about reviewers *Reviewer History Macro* has a statistically significant effect on the perceived helpfulness of reviews of digital cameras and DVDs, but interestingly, the directional impact is quite mixed across these two categories. In sum, our results provide support for Hypotheses 2a to 2d.

Note that the within R-squared values of our models range between 0.02 and 0.08 across the four product categories. This is because these R-squared values are for the "within" (differenced) fixed effect estimator that estimates this regression by differencing out the average values across product-sellers. The R-squared reported is obtained by only fitting a mean deviated model where the effects of the groups (all of the dummy variables for the products) are assumed to be fixed quantities. So, all of the effects for the groups are simply subtracted out of the model and no attempt is made to quantify their overall effect on the fit of the model. This means that the calculated "within" R-squared values do not take into account the explanatory power of the fixed effects. If we estimate the fixed effects instead of differencing them out, the measured R-squared would be much higher. However this becomes computationally unattractive. This is consistent with prior work([5]).[11]

Our econometric analyses imply that we can quickly estimate the helpfulness of a review by performing an automatic stylistic analysis in terms of subjectivity, readability and linguistic correctness. Hence, we can immediately identify reviews that are likely to have a 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. However, a key limitation of this existing feature is that it because it relies on a sufficient number of people to vote on reviews, it requires a long time to elapse before identifying a helpful review.

5 Predictive Modeling

The explanatory study that we described above revealed what factors influence the helpfulness and impact of a review. In this section, we switch from explanatory modeling to *predictive* modeling. In other words, the main goal now is not to explain which factors affect helpfulness and impact, but to examine whether, given an existing review, how well can we predict the helpfulness and economic impact of an *unseen* review, i.e., of a review that was not included in the data used to train the predictive model.

[11] As before, we compared our work with the model used in [5] who examined the drivers of review helpfulness but without incorporating the impact of review text variables and without incorporating the reviewer history-level variables. Our model increases R-squared by 5% for audio-video products, 8% for digital cameras and 5% for DVDs.

5.1 Predicting Helpfulness

The *Helpfulness* of each review in our data set is defined by the votes of the peer customers, who decide whether a review is helpful or not. In our predictive framework, we could use a regression model, as in Section 4, or use a classification approach and build a binary prediction model that classifies a review as helpful or not. We attempted both approaches and the results were similar. Since we have already described a regression framework in Section 4, we now focus instead on a binary prediction model for brevity. In the rest of the section, we first describe our methodology for converting the continuous helpfulness variable into binary. Then we describe the results of our experiments, using various machine learning approaches.

5.2 Converting Continuous Helpfulness to Binary

Converting the continuous variable *Helpfulness* into a binary one is, in principle, a straightforward process. Since *Helpfulness* goes from 0 to 1, we can simply select a threshold τ, and mark all reviews that have $helpfulness \geq \tau$ as *helpful* and the others as *not helpful*. However, selecting the proper value for the threshold τ is slightly trickier. What is the "best" value of τ for separating the "helpful" from the "not helpful" reviews? Setting τ too high would mean that helpful reviews would be classified as not helpful, and setting τ too low would have the opposite effect.

In order to select a good value for τ, we used two human coders do a content analysis on a sample of 1,000 reviews. The reviews were randomly chosen for each category. The main aim was to analyze whether the review was informative. For this, we asked the the coders to read each review and answer the question "*Is the review informative or not?*" The coders did not have access to the helpful and total votes that were casted for the review, but could see the star rating and the product that the review was referring to. We measured the inter-rater agreement across the two coders, using the kappa statistic. The analysis showed a substantial agreement, with $\kappa = 0.739$.

Our next step was to identify the optimal threshold (in terms of percentage of helpful votes) that separates 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 "helpful". Otherwise, the review is classified as "not helpful" and this decision achieves a good balance between false positive errors and false negative errors.

Our analysis is presented in Figure 1. On the x-axis, we have the decision threshold τ, which is the percentage of useful votes out of all the votes received by a given review. Each review is marked as "useful" or "not-useful" by our coders, independently of the peer votes actually posted on Amazon.com. Based on the coder's classification, we compute the: (a) percentage of useful reviews that have an Amazon helpfulness rating below τ, (b) percentage of not-useful reviews that have an Amazon helpfulness rating above τ. (These values are essentially the error rates for the two classes if we set the decision threshold at τ.) Furthermore, by considering the "useful" class as the positive class, we compute the precision and recall metrics. We can see that if we set the separation threshold at 0.6, then the error rate in the classification is minimized. For this reason, we pick 0.6 as the threshold of separating the reviews as "useful" or not. In other words, if more than 60% of the votes indicate that the review is helpful, then we classify a review as "useful." Otherwise, the review is classified as "non-useful."

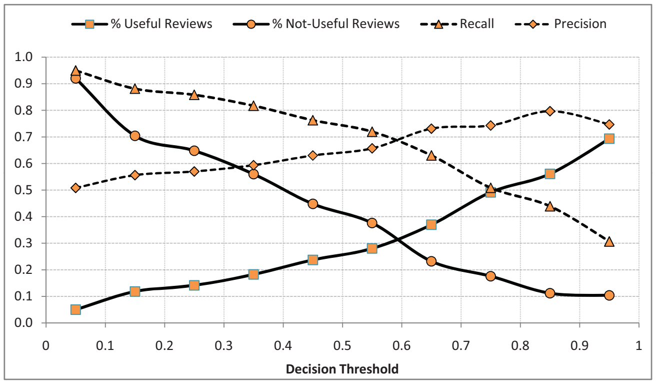

Fig. 1. Picking a decision threshold that minimizes error rates for converting the continuous helpfulness variable into a binary one.

5.3 Building the Predictive Model

Once we are able to separate the reviews into two classes, we can then use any supervised learning technique to learn a model that classifies an unseen review as helpful or not.

We experimented with Support Vector Machines [51] and Random Forests [52]. Support Vector Machines have been reported to work well in the past for the problem of predicting review helpfulness. However, in all our experiments, SVM's consistently performed worse than Random Forests, for both our techniques and for the existing baselines, such as for the algorithm of Zhang and Varadarajan [23] that we used as a baseline for comparison. Furthermore, training time was significantly higher for SVM's compared to Random Forests. This empirical finding is consistent with recent comparative experiments [53], [54] that indicate that Random Forests are robust and perform better than SVM's for a variety of learning tasks. Therefore, in this experimental section we report only the results that we obtained using Random Forests.

In our experiments with Random Forests, we use 20 trees and we generate a different classifier for each product category. Our evaluation results are based on stratified 10-fold cross validation and we use as evaluation metrics the classification accuracy and the area under the ROC curve (AUC).

Using All Available Features: In our first experiment, we used all the features that we had available to build the classifiers. The resulting performance of the classifiers was quite high, as seen in Table 7.

One interesting result is the relatively lower predictive performance of the classifier that we constructed for the *DVD* data set. This can be explained by the nature of the goods: DVDs are *experience goods* whose quality is difficult to estimate in advance but can be ascertained after consumption. In contrast, digital cameras and audio & video are *search goods*,

Data Set	**Accuracy**	**AUC**
DVD	78.79%	0.73
Audio & Video	87.57%	0.94
Digital Cameras	87.68%	0.94

TABLE 7
Accuracy and Area under the ROC curve for the Helpfulness Classifiers

Data Set	Features	**Accuracy**	**AUC**
DVD	Baseline [23]	65.25%	0.58
	Reviewer	78.19%	0.71
	Subjectivity	77.95%	0.72
	Readability	77.23%	0.69
	Reviewer + Subjectivity	78.72%	0.73
	Reviewer + Readability	78.09%	0.72
	Subjectivity + Readability	78.14%	0.74
Audio & Video	Baseline [23]	69.45%	0.76
	Reviewer	83.07%	0.89
	Subjectivity	85.42%	0.91
	Readability	86.68%	0.94
	Reviewer + Subjectivity	86.82%	0.94
	Reviewer + Readability	87.11%	0.94
	Subjectivity + Readability	85.64%	0.93
Digital Cameras	Baseline [23]	70.26%	0.71
	Reviewer	82.99%	0.84
	Subjectivity	84.40%	0.86
	Readability	87.68%	0.93
	Reviewer + Subjectivity	86.47%	0.92
	Reviewer + Readability	89.04%	0.94
	Subjectivity + Readability	84.80%	0.90

TABLE 8
Accuracy and Area under the ROC curve for the Helpfulness Classifiers

i.e., products with features and characteristics easily evaluated before purchase. Therefore the notion of *helpfulness* is more subjective for experience goods, as what constitutes a helpful review for one customer is not necessarily helpful for another. This contrasts with the search goods, in which a good review is one that allows customers to evaluate better, before the purchase, the quality of the underlying good.

Going beyond the aggregate results, we examined what kinds of reviews have helpfulness scores that are most difficult to predict. Interestingly, we observed a high correlation of classification error with the distribution of the underlying review ratings. Reviews for products that have received widely fluctuating reviews, also have reviews of widely fluctuating helpfulness. However, the different helpfulness scores do not necessarily correspond well with reviews of different "inherent" quality. Rather, in such cases, customers tend to vote *not* on the merits of the review per se, but rather to convey their approval or disapproval of the rating of the review. In such cases, an otherwise detailed and helpful review, may receive a bad helpfulness score. This effect was more pronounced in the DVD category, but also appeared in the digital camera and in the audio & video category. Even though this observation did not help us improve the predictive accuracy of our model, it is a good heuristic for estimating a-priori the predictive accuracy of our models for reviews of such products.

Examining the Predictive Power of Features: The next step was to examine what is the power of the different features that we have generated. As can be seen from Table 1, we have three broad feature categories: (i) *reviewer features* that include both *reviewer history* and *reviewer characteristics*, (ii) *review subjectivity features*, and (iii) *review readability features*. To examine their importance, we built classifiers using only subsets of the features. As a comparison, we also list the results that we got by using the features used by Zhang and Varadarajan [23], which we refer to as "Baseline." We evaluated each classifier in the same way as above, using stratified 10-fold cross validation, and reporting the accuracy and the area under the ROC curve.

The first result that we observed is that our techniques clearly outperformed the existing baseline from [23]: the increased predictive performance of our models was rather anticipated given the difference in R^2 values in the explanatory regression models. The R^2 values for the regressions in [23] were around 0.3-0.4, while our explanatory econometric models achieved R^2 values in the 0.7-0.9 area. This difference in performance in the training set was also visible in the predictive performance of the models.

Another interesting result is that using any of the feature sets resulted in only a modest decrease in performance compared to the case of using all available features. To explore further this puzzling result, we conducted an additional experiment: We examined whether we could predict the value of the features in one set, using the features from the other two feature sets (e.g., predict *review subjectivity* using the *reviewer-related* and *review readability* features). We conducted the tests for all combinations. Surprisingly, the results indicated that the three feature sets are interchangeable. In other words, the information in the *review readability* and *review subjectivity* set is enough to predict the value of variables in the *reviewer* set, and vice versa. Reviewers who have historical generated helpful reviews, tend to post reviews of specific readability levels, and with specific subjectivity mixtures in the reviews. Even though this may seem counterintuitive at first, it simply indicates that there is correlation between these variables, not causation. Identifying causality is rather difficult and is beyond the scope of this paper. What is of interest in our case is that the three feature sets are roughly equivalent in terms of predictive power.

5.4 Predicting Impact on Sales

Our analysis so far, indicated that we can successfully predict whether a review is going to be rated as helpful by the peer customers or not. The next task that we wanted to examine was whether we can predict the impact of a review on sales.

Specifically, we examine whether the review characteristics can be used to predict whether the (comparative) sales of a product will go up or down after a review is published. So, we examine whether the difference

$$SalesRank_{t(r)+T} - SalesRank_{t(r)}$$

where $t(r)$ is the time the review is posted, is positive or negative. Since the effect of a review is not immediate, we examine variants of the problem for $T = 1$, $T = 3$, $T = 7$, $T = 14$ (in days). By having different time intervals, we wanted to examine how far in the future we can extend our prediction and still get reasonable results.

Data Set	Features	**Accuracy**	**AUC**
DVD	$T = 1$	82.28%	0.91
	$T = 3$	82.28%	0.91
	$T = 7$	83.06%	0.92
	$T = 14$	84.35%	0.92
Audio & Video	$T = 1$	80.60%	0.91
	$T = 3$	81.28%	0.91
	$T = 7$	81.75%	0.92
	$T = 14$	83.24%	0.93
Digital Cameras	$T = 1$	73.98%	0.82
	$T = 3$	74.43%	0.86
	$T = 7$	77.93%	0.87
	$T = 14$	79.47%	0.89

TABLE 9
Accuracy and Area under the ROC curve for the Sales Impact Classifiers

As we can see from the results in Table 9, the prediction accuracy is high, demonstrating that we can predict the direction of sales given the review information. While it is hard, at this point, to claim causality (it is unclear whether the reviews influence sales, or whether the reviews are just a manifestation of the underlying sales trend), it is definitely possible to show a strong correlation between the two. We also observed that the predictive power increases slightly as T increases, indicating that the influence of a review is not immediate.

We also performed experiments with subsets of features as in the case of helpfulness. The results were very similar to the case of helpfulness: training models with subsets of the features results in similar predictive power. Given the helpfulness results, which we discussed above, this should not be a surprise.

6 Conclusions and Future Work

In this paper, we build on our previous work [20], [21] by expanding our data to include multiple product categories and multiple textual features such as different readability metrics, information about the reviewer history, different features of reviewer disclosure and so on. The present paper is unique in looking at how *subjectivity* levels, *readability* and *spelling errors* in the text of reviews affect product sales and the perceived helpfulness of these reviews.

Our key results from both the econometric regressions and predictive models can be summarized as follows:

- Based on Hypothesis 1a, we find that an increase in the average subjectivity of reviews is associated with an increase in sales for products. Further, a decrease in the deviation of the probability of subjective comments is associated with an increase in product sales. This means that reviews that have a mixture of objective, and highly subjective sentences are negatively associated with product sales, compared to reviews that tend to include only subjective or only objective information.
- Based on Hypothesis 1b, we find that for some products like digital cameras reviews that have higher Readability scores are associated with higher sales. Based on Hypothesis 1c, we find that an increase in the proportion of spelling mistakes in the content of the reviews decreases product sales for some experience products like DVDs whose quality can be assessed only after purchase. However, for search products such as audio-video players and digital cameras, the proportion of spelling errors in reviews does not have a statistically significant impact on sales.
- Further, reviews with that rate products *negatively* can be associated with *increased product sales* when the review text is *informative and detailed*. This is likely to occur when the reviewer clearly outlines the pros and cons of the product, thereby providing sufficient information to the consumer to make a purchase.
- Based on Hypothesis 2a, we find that in general, reviews which tend to include a mixture of subjective and objective elements are considered more informative (or helpful) by the users. In terms of subjectivity and its effect on helpfulness, we observe that for feature-based goods, such as electronics, users prefer reviews that contain mainly objective information with only a few subjective sentences and rate those higher. In other words, users prefer reviews that 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 DVDs, the marginally significant coefficient on subjectivity suggests that while users do prefer to see a brief description of the “objective” elements of the good (e.g., the plot), they do expect to see a personalized, highly sentimental positioning, describing aspects of the movie that are not captured by the product description provided by the producers.
- Based on Hypothesis 2b through 2d, we find that an increase in the readability of reviews has a positive and statistically impact on review helpfulness while an increase in the proportion of spelling errors has a negative and statistically significant impact on review helpfulness for audio-video products and DVDs. While the past historical information about reviewers has a statistically significant effect on the perceived helpfulness of reviews, interestingly enough, the directional impact is quite mixed across different product categories.
- Using Random Forest classifiers, we find that for experience goods like DVDs, the classifiers have a lower performance while predicting the helpfulness of reviews, compared to that for search goods like electronics products. Furthermore, we observe a high correlation of classification error with the distribution of the underlying review ratings. Reviews for products that have received widely fluctuating ratings, also have reviews with widely fluctuating helpfulness votes. In particular, we found evidence that highly detailed and readable reviews can have low helpfulness votes in cases when users tend to vote negatively not because they disapprove of the review quality (extent of helpfulness) but rather to convey their disapproval of the rating provided by the reviewer for that review.
- Finally, we examined the relative importance of the three broad feature categories: ‘reviewer-related’ features, ‘review subjectivity’ features, and ‘review readability’ features. We found that using any of the three feature sets resulted in a statistically equivalent performance as in the case of using all available features. Further, we find that

the three feature sets are interchangeable. In other words, the information in the 'readability' and 'subjectivity' set is sufficient to predict the value of variables in the 'reviewer' set, and vice versa. Experiments with classifiers for predicting sales yield similar results in terms of the inter-changeability of the three broad feature sets.

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.

While we have taken a first step examining the economic value of textual content in word-of-mouth forums, we acknowledge that our approach has several limitations, many of which are borne by the nature of the data itself. Some of the variables in our data are proxies for the actual measure that one would need for more advanced empirical modeling. For example, we use sales rank as a proxy for demand in accordance with prior work. Future work can look at real demand data. Our sample is also restricted in that our analysis focuses on the sales at one e-commerce retailer. The actual magnitude of the impact of textual information on sales may be different for a different retailer. Additional work in other on-line contexts will be needed to evaluate whether review text information has similar explanatory power that are similar to those we have obtained.

There are many other interesting directions to follow when analyzing online reviews. For example, in our work, we analyzed each review independently of the other, existing reviews. A recent stream of research indicates that the helpfulness of a review is also a function of the other submitted reviews [55] and that temporal dynamics can play a role in the perceived helpfulness of a review [25], [56], [57] (e.g., early reviews, everything else being equal, get higher helpfulness scores). Furthermore, the helpfulness of a review may be influenced by the way that reviews are presented to different types of users [58] and by the context in which a user evaluates a given review [59].

Overall, we consider this work a significant first step in understanding the factors that affect the perceived quality and economic impact of reviews and believe that there are many interesting problems that need to be addressed in this area.

ACKNOWLEDGMENTS

The authors would like to thank Rong Zheng, Ashley Tyrrel and Leslie Culpepper for assistance in data collection. We want to thank the participants in WITS 2006, SCECR 2007 and ICEC 2007 and the reviewers for very helpful comments. This work was partially supported by NET Institute (http://www.NETinst.org), a Microsoft Live Labs Search Award, a Microsoft Virtual Earth Award, New York University Research Challenge Fund grants N-6011 and R-8637, and by NSF grants IIS-0643847, IIS-0643846, and DUE-0911982. 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] N. Hu, P. A. Pavlou, and J. Zhang, "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.

[2] C. Dellarocas, N. F. Awad, and X. M. Zhang, "Exploring the value of online product ratings in revenue forecasting: The case of motion pictures," 2007, working Paper, Robert H. Smith School Research Paper.

[3] J. A. Chevalier and D. Mayzlin, "The effect of word of mouth on sales: Online book reviews," *Journal of Marketing Research*, vol. 43, no. 3, pp. 345–354, Aug. 2006.

[4] D. Reinstein and C. M. Snyder, "The influence of expert reviews on consumer demand for experience goods: A case study of movie critics," *Journal of Industrial Economics*, vol. 53, no. 1, pp. 27–51, Mar. 2005.

[5] C. Forman, A. Ghose, and B. Wiesenfeld, "Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets," *Information Systems Research*, vol. 19, no. 3, Sep. 2008.

[6] Y. Liu, "Word of mouth for movies: Its dynamics and impact on box office revenue," *Journal of Marketing*, vol. 70, no. 3, pp. 74–89, Jul. 2006.

[7] W. Duan, B. Gu, , and A. B. Whinston, "The dynamics of online word-of-mouth and product sales. An empirical investigation of the movie industry," *Journal of Retailing*, vol. 84, no. 2, pp. 233–242, 2008.

[8] R. G. Hass, "Effects of source characteristics on cognitive responses and persuasion," in *Cognitive responses in persuasion*, R. E. Petty, T. M. Ostrom, and T. C. Brock, Eds. Lawrence Erlbaum Associates, 1981, pp. 1–18.

[9] S. Chaiken, "Heuristic versus systematic information processing and the use of source versus message cues in persuasion," *Journal of Personality and Social Psychology*, vol. 39, no. 5, pp. 752–766, 1980.

[10] ——, "The heuristic model of persuasion," in *Social Influence: The Ontario Symposium, Volume 5*, M. P. Zanna, J. M. Olson, and C. P. Herman, Eds. Lawrence Erlbaum Associates, 1987, pp. 3–39.

[11] J. J. Brown and P. H. Reingen, "Social ties and word-of-mouth referral behavior," *Journal of Consumer Research*, vol. 14, no. 3, pp. 350–362, Dec. 1987.

[12] R. Spears and M. Lea, "Social influence and the influence of the 'social' in computer-mediated communication," in *Contexts of Computer-mediated Communication*, M. Lea, Ed. Harvester Wheatsheaf, Jun. 1992, pp. 30–65.

[13] S. L. Jarvenpaa and D. E. Leidner, "Communication and trust in global virtual teams," *Journal of Interactive Marketing*, vol. 10, no. 6, pp. 791–815, Nov.-Dec. 1999.

[14] K. Y. A. McKenna and J. A. Bargh, "Causes and consequences of social interaction on the internet: A conceptual framework," *Media Psychology*, vol. 1, no. 3, pp. 249–269, Sep. 1999.

[15] U. M. Dholakia, R. P. Bagozzi, and L. K. Pearo, "A social influence model of consumer participation in network- and small-group-based virtual communities," *International Journal of Research in Marketing*, vol. 21, no. 3, pp. 241–263, Sep. 2004.

[16] T. Hennig-Thurau, K. P. Gwinner, G. Walsh, and D. D. Gremler, "Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the internet?" *Journal of Interactive Marketing*, vol. 18, no. 1, pp. 38–52, 2004.

[17] M. Ma and R. Agarwal, "Through a glass darkly: Information technology design, identity verification, and knowledge contribution in online communities," *Information Systems Research*, vol. 18, no. 1, pp. 42–67, Mar. 2007.

[18] A. Ghose, P. G. Ipeirotis, and A. Sundararajan, "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, pp. 416–423.

[19] N. Archak, A. Ghose, and P. G. Ipeirotis, "Show me the money! Deriving the pricing power of product features by mining consumer reviews," in *Proceedings of the Twelveth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2007)*, 2007, pp. 56–65.

[20] A. Ghose and P. G. Ipeirotis, "Designing ranking systems for consumer reviews: The impact of review subjectivity on product sales and review quality," in *Workshop on Information Technology and Systems*, 2006.

[21] ——, "Designing novel review ranking systems: Predicting the usefulness and impact of reviews," in *Proceedings of the 9th International Conference on Electronic Commerce (ICEC 2007)*, 2007, pp. 303–310.

[22] ——, "Estimating the socio-economic impact of product reviews: Mining text and reviewer characteristics," Center for Digital Economy Research, New York University, Tech. Rep. CeDER-08-06, Sep. 2008.

[23] Z. Zhang and B. Varadarajan, "Utility scoring of product reviews," in *Proceedings of the 2006 ACM Conference on Information and Knowledge Management (CIKM 2006)*, 2006, pp. 51–57.

[24] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti, "Automatically assessing review helpfulness," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, 2006, pp. 423–430.

[25] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou, "Low-quality product review detection in opinion summarization," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007, pp. 334–342.

[26] J. Otterbacher, “Helpfulness in online communities: A measure of message quality,” in *CHI ’09: Proceedings of the 27th international conference on Human factors in computing systems*, 2009, pp. 955–964.

[27] Y. Liu, X. Huang, A. An, and X. Yu, “Modeling and predicting the helpfulness of online reviews,” in *Eighth IEEE International Conference on Data Mining (ICDM’08)*, 2008, pp. 443–452.

[28] M. Weimer, I. Gurevych, and M. Mühlhäuser, “Automatically assessing the post quality in online discussions on software,” in *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, 2007, pp. 125–128.

[29] M. Weimer and I. Gurevych, “Predicting the perceived quality of web forum posts,” in *Proceedings of the 2007 Conference on Recent Advances in Natural Language Processing, RANLP 2007*, 2007.

[30] J. Jeon, W. B. Croft, J. H. Lee, and S. Park, “A framework to predict the quality of answers with non-textual features,” in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2006*, 2006, pp. 228–235.

[31] L. Hoang, J.-T. Lee, Y.-I. Song, and H.-C. Rim, “A model for evaluating the quality of user-created documents,” in *Proceedings of the 4th Asia Infomation Retrieval Symposium (AIRS 2008)*, 2008, pp. 496–501.

[32] Y. Y. Hao, Y. J. Li, and P. Zou, “Why some online product reviews have no usefulness rating?” in *Pacific Asia Conference on Information Systems (PACIS 2009)*, 2009.

[33] O. Tsur and A. Rappoport, “Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews,” in *Third International AAAI Conference on Weblogs and Social Media (ICWSM 2009)*, 2009.

[34] A. Ghose and A. Sundararajan, “Evaluating pricing strategy using e-commerce data: Evidence and estimation challenges,” *Statistical Science*, vol. 21, no. 2, pp. 131–142, May 2006.

[35] V. Hatzivassiloglou and K. R. McKeown, “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.

[36] M. Hu and B. Liu, “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.

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

[38] P. D. Turney, “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.

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

[40] K. Dave, S. Lawrence, and D. M. Pennock, “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.

[41] B. Pang and L. Lee, “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.

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

[43] B. Pang and L. Lee, “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.

[44] T. Wilson, J. Wiebe, and R. Hwa, “Recognizing strong and weak opinion clauses,” *Computational Intelligence*, vol. 22, no. 2, pp. 73–99, May 2006.

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

[46] B. Snyder and R. Barzilay, “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.

[47] S. White, “The 2003 national assessment of adult literacy (NAAL),” Center for Education Statistics (NCES), Institute of Education Sciences, U.S. Department of Education, Tech. Rep. NCES 2003495rev, Mar. 2003, also available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2003495rev.

[48] W. H. DuBay, *The Principles of Readability*. Impact Information, 2004, available at http://www.nald.ca/library/research/readab/readab.pdf.

[49] J. A. Chevalier and A. Goolsbee, “Measuring prices and price competition online: Amazon.com and BarnesandNoble.com,” *Quantitative Marketing and Economics*, vol. 1, no. 2, pp. 203–222, 2003.

[50] J. M. Wooldridge, *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, 2001.

[51] C. J. C. Burges, “A tutorial on support vector machines for pattern recognition,” *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, Jun. 1998.

[52] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[53] R. Caruana and A. Niculescu-Mizil, “An empirical comparison of supervised learning algorithms,” in *Proceedings of the 23rd International Conference on Machine Learning (ICML 2006)*, 2006, pp. 161–168.

[54] R. Caruana, N. Karampatziakis, and A. Yessenalina, “An empirical evaluation of supervised learning in high dimensions,” in *Proceedings of the 25th International Conference on Machine Learning (ICML 2008)*, 2008.

[55] C. Danescu-Niculescu-Mizil, G. Kossinets, J. Kleinberg, and L. Lee, “How opinions are received by online communities: A case study on Amazon.com helpfulness votes,” in *Proceedings of the 18th international conference on World Wide Web (WWW 2009)*, 2009, pp. 141–150.

[56] W. Shen, “Essays on online reviews: The strategic behaviors of online reviewers to compete for attention, and the temporal pattern of online reviews,” 2008, ph.D. Proposal, Krannert Graduate School of Management Purdue University.

[57] Q. Miao, Q. Li, and R. Dai, “Amazing: A sentiment mining and retrieval system,” *Expert Systems with Applications*, vol. 36, no. 3, Part 2, pp. 7192–7198, 2009.

[58] P. Victor, C. Cornelis, M. De Cock, and A. Teredesai, “Trust- and distrust-based recommendations for controversial reviews,” in *Proceedings of the WebSci’09: Society On-Line*, 2009.

[59] S. A. Yahia, A. Z. Broder, and A. Galland, “Reviewing the reviewers: Characterizing biases and competencies using socially meaningful attributes,” in *AAAI Spring Symposium*, 2008.

Anindya Ghose is an Associate Professor of Information, Operations, and Management Sciences and Robert L. and Dale Atkins Rosen Faculty Fellow at New York University’s Leonard N. Stern School of Business. He is an expert in building econometric models to quantify the economic value from user-generated content in social media; estimating the impact of search engine advertising; modeling consumer behavior in mobile media and mobile Internet; and measuring the welfare impact of the Internet. He has worked on product reviews, reputation and rating systems, sponsored search advertising, mobile Internet, mobile social networks, and online used-good markets. His research has received best paper awards and nominations at leading journals and conferences such as ICIS, WITS and ISR. In 2007, he received the prestigious NSF CAREER Award. He is also a winner of a ACM SIGMIS Doctoral Dissertation Award, a Microsoft Live Labs Award, a Microsoft Virtual Earth Award, a Marketing Science Institute grant, several WIMI (Wharton Interactive Media Initiative) awards, a NSF SFS award, a NSF IGERT award, and a Google-WPP Marketing Research Award. He serves as an Associate Editor of Management Science and Information Systems Research.

Panagiotis G. Ipeirotis is an Associate Professor and George A. Kellner Faculty Fellow at the Department of Information, Operations, and Management Sciences at Leonard N. Stern School of Business of New York University. His recent research interests focus on crowdsourcing and on mining user-generated content on the Internet. He received his Ph.D. degree in Computer Science from Columbia University in 2004, with distinction. He has received two ”Best Paper” awards (IEEE ICDE 2005, ACM SIGMOD 2006), two ”Best Paper Runner Up” awards (JCDL 2002, ACM KDD 2008), and is also a recipient of a CAREER award from the National Science Foundation.