MANAGEMENT SCIENCE
Vol. 57, No. 8, August 2011, pp. 1485–1509
ISSN 0025-1909 | EISSN 1526-5501 | 11 | 5708 | 1485

informs

DOI 10.1287/mnsc.1110.1370

Deriving the Pricing Power of Product Features by Mining Consumer Reviews

Nikolay Archak, Anindya Ghose, Panagiotis G. Ipeirotis
Leonard Stern School of Business, New York University, New York, New York 10012
{narchak@stern.nyu.edu, aghose@stern.nyu.edu, panos@stern.nyu.edu}

Increasingly, user-generated product reviews serve as a valuable source of information for customers making product choices online. The existing literature typically incorporates the impact of product reviews on sales based on numeric variables representing the valence and volume of reviews. In this paper, we posit that the information embedded in product reviews cannot be captured by a single scalar value. Rather, we argue that product reviews are multifaceted, and hence the textual content of product reviews is an important determinant of consumers' choices, over and above the valence and volume of reviews. To demonstrate this, we use text mining to incorporate review text in a consumer choice model by decomposing textual reviews into segments describing different product features. We estimate our model based on a unique data set from Amazon containing sales data and consumer review data for two different groups of products (digital cameras and camcorders) over a 15-month period. We alleviate the problems of data sparsity and of omitted variables by providing two experimental techniques: clustering rare textual opinions based on pointwise mutual information and using externally imposed review semantics. This paper demonstrates how textual data can be used to learn consumers' relative preferences for different product features and also how text can be used for predictive modeling of future changes in sales.

Key words: Bayesian learning; consumer reviews; discrete choice; electronic commerce; electronic markets; opinion mining; sentiment analysis; user-generated content; text mining; econometrics
History: Received November 13, 2008; accepted February 23, 2011, by Ramayya Krishnan, information systems. Published online in *Articles in Advance* June 30, 2011.

1. Introduction

The growing pervasiveness of the Internet has changed the way that consumers shop for goods. Whereas in a "brick-and-mortar" store visitors can usually test and evaluate products before making purchase decisions, in an online store their ability to directly assess product value is significantly more limited. Online shoppers increasingly rely on alternative sources of information such as "word of mouth" in general, and user-generated product reviews in particular. In fact, some researchers have established that user-generated product information on the Internet attracts more interest than vendor information among consumers (Bickart and Schindler 2001). In contrast to product descriptions provided by vendors, consumer reviews are, by construction, more user oriented. In a review, customers describe the product in terms of different usage scenarios and evaluate it from the user's perspective (Chen and Xie 2008). Despite the subjectivity of consumer evaluations in the reviews, such evaluations are often considered more credible and trustworthy by customers than traditional sources of information (Bickart and Schindler 2001).

The hypothesis that product reviews affect product sales has received strong support in prior empirical studies (Godes and Mayzlin 2004, Duan et al. 2005, Chevalier and Mayzlin 2006, Liu 2006, Dellarocas et al. 2007, Forman et al. 2008, Ghose and Ipeirotis 2010, Ghose et al. 2011). However, these studies have only used the numeric review ratings (e.g., the number of stars) and the volume of reviews in their empirical analysis, without formally incorporating the information contained in the text of the reviews. To the best of our knowledge, only a handful of empirical studies have formally tested whether the textual information embedded in online user-generated content can have an economic impact. Ghose et al. (2007) estimate the impact of buyer textual feedback on price premiums charged by sellers in online second-hand markets. Eliashberg et al. (2007) combine natural-language-processing techniques and statistical learning methods to forecast the return on investment for a movie, using shallow textual features from movie scripts. Netzer et al. (2011) combine text mining and semantic network analysis to understand the brand associative network and the implied market structure. Decker and Trusov (2010) use text mining to estimate the relative effect of product attributes and brand names on the overall evaluation of the products. But none of these studies focus on estimating the impact of user-generated product reviews in influencing product sales beyond the effect of numeric

review ratings, which is one of the key research objectives of this paper. The papers closest to this paper are those by Ghose and Ipeirotis (2010) and Ghose et al. (2011), who explore multiple aspects of review text, such as lexical, grammatical, semantic, and stylistic levels to identify important text-based features and study their impact on review helpfulness (Ghose and Ipeirotis 2010) and product sales (Ghose and Ipeirotis 2010, Ghose et al. 2011). However, they do not focus on examining the economic impact of different product attributes and opinions on product sales.

There is a potential issue with using only numeric ratings as being representative of the information contained in product reviews. By compressing a complex review to a single number, we implicitly assume that the product quality is one-dimensional, whereas economic theory (see, for example, Rosen 1974) tells us that products have multiple attributes and different attributes can have different levels of importance to consumers. Tastes for product attributes tend to vary across individuals. Thus, unless the person reading a review has exactly the same preferences as the person who wrote the review, a single number, like an average product rating, might not be sufficient for the reader to extract all information relevant to the purchase decision.

Moreover, it has been shown that idiosyncratic preferences of early buyers can affect long-term consumer purchase behavior and that rating can have a self-selection bias (Li and Hitt 2008). Consequently, Li and Hitt (2008) suggest that consumer-generated product reviews may not be an unbiased indication of unobserved product quality. Furthermore, recent work has shown that the distribution of an overwhelming majority of reviews posted in online markets is bimodal (Hu et al. 2008). In such situations, the average numerical star rating assigned to a product may not convey a lot of information to a prospective buyer. Instead, the reader has to read the actual reviews to examine which of the positive and which of the negative attributes of a product are of interest. Furthermore, there may be extra information in the text because of the discreteness problem: Reviews are allowed to be rated only as an integer from 1 to 5. However, some "4" reviews read like "3" reviews, whereas others read like "5" reviews. Therefore, our second research objective in this paper is to analyze the extent to which product reviews can help us learn consumer preferences for different product attributes and how consumers make trade-offs between different attributes.

The key challenge is in bridging the gap between the essentially textual and qualitative nature of review content and the quantitative nature of discrete choice models. Any successful attempt to address this challenge necessitates an answer to the following questions.

1. How can we identify which product attributes are evaluated in a product review?
2. How can we extract opinions about the product attributes expressed in a product review?
3. How can we model the economic impact of these extracted opinions?

With the rapid growth and popularity of user-generated content on the Web, a new area of research applying text-mining techniques to content analysis of product reviews has emerged. The first stream of this research has focused on sentiment analysis of product reviews. The earliest work in this area was targeted primarily at evaluating the *polarity* of a review. Reviews were classified as positive or negative based on the occurrences of specific sentiment phrases (Das and Chen 2007, Hu and Liu 2004). More recent work has suggested that sentiment classification of consumer reviews is complicated, because consumers may provide a mixed review by praising some aspects of a product but criticizing other. This stimulated additional research on identifying product features in reviews (Hu and Liu 2004, Ghani et al. 2006). Automated extraction of product attributes has also received attention in the recent marketing literature. In particular, Lee and Bradlow (2007) present an automatic procedure for obtaining conjoint attributes and levels through the analysis of Epinions reviews that list the explicit pros and cons of a product. Pang and Lee (2008) offer an excellent and comprehensive survey of the research in the field of sentiment analysis.

So, how does this paper contribute to prior research? Prior work in text mining does not reliably capture the pragmatic meaning of the customer evaluations; in particular, the existing approaches do not provide *quantitative* evaluations of product features. In most cases, the evaluation of a product feature is done in a binary manner (positive or negative). It is also possible to use a counting scale to compute the number of positive and negative opinion sentences for a particular feature; opinion counts can later be used for the feature-based comparison of two products (Liu et al. 2005). Such a comparison tool is undoubtedly useful for consumers using an online shopping environment. Unfortunately, this technique ignores the strength of the evaluations and does not demonstrate the importance of the product feature in the consumers' choice process. Is *"good battery life"* more important for a digital camera than a *"small size"*? If so, then how important is it in influencing the purchase decision? Although questions of this nature might seem fuzzy, they can gain meaning if evaluated in the economic context surrounding consumer reviews and sales.

In sum, our paper aims to infer the economic impact of user-generated product reviews by identifying the weight that consumers put on individual evaluations and product features, and estimating

the overall impact of review text on sales. We do so by using both econometric and predictive modeling methods. Our paper can be considered an extension of the prior work of Chevalier and Mayzlin (2006) that incorporates textual consumer opinions directly in a reduced-form equation for product demand. A justification of our empirical modeling approach based on a theoretical model of multiattribute choice under uncertainty is described in this paper, with details given in the appendix. We compare estimation results from the inferred polarity model with a model in which the polarity is imposed ex ante from a predefined ontology. We alleviate the problems of data sparsity and of omitted variables by providing two experimental techniques: clustering rare textual opinions based on pointwise mutual information and using externally imposed review semantics.

For estimation, we use a 15-month panel of product sales and reviews of digital cameras and camcorders retrieved from Amazon. To properly capture both longitudinal and cross-sectional properties of our data set, we apply generalized method of moments (GMM)-based dynamic panel data estimators. We additionally consider a purely predictive problem of forecasting product sales based on textual review contents. Results demonstrate that our text-mining approach delivers an explicit improvement in the out-of-sample forecasting performance.

The econometric modeling approach we adopt can be compared to the *hedonic regressions* that are commonly used in econometrics to identify the weight of individual features in determining the overall price of a product. However, instead of studying the relationship between the fixed and objective product qualities and the product price, we study the relationship between *beliefs* about features that are either not directly measurable or are qualitative in nature and product *demand*.

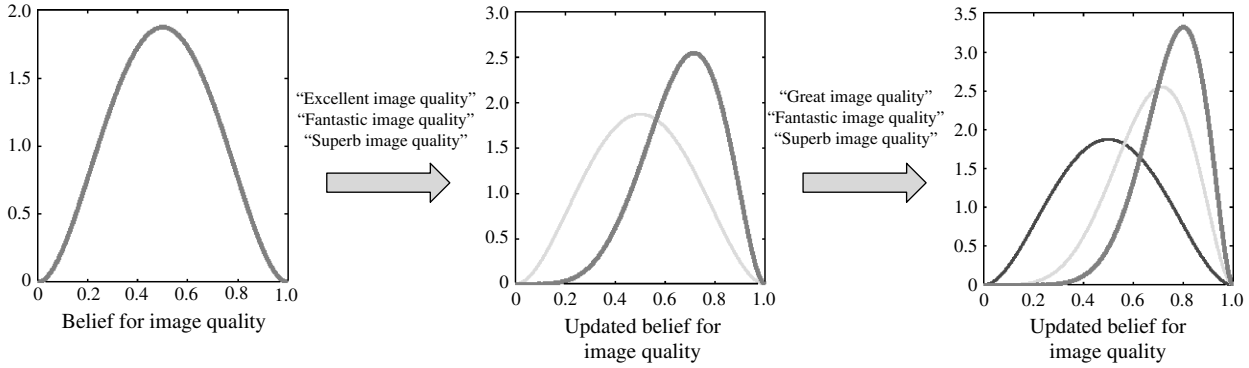
Our approach also differs from classic discrete choice models, such as BLP (Berry et al. 1995). Similar to the BLP model, we study substitution patterns. However, the nature of the patterns that we capture is somewhat different. In a typical discrete choice model, a consumer can switch from one product to another product either when a new product is introduced in the market or when some attribute of an existing product changes. Because most of the product attributes generally do not change after the introduction of the product, substitutions happen mostly because of new product introductions and variation in prices of the existing products. There is generally no uncertainty about product qualities in standard discrete choice models. To the contrary, we think of online consumers as having certain beliefs about features of the products offered. As new product reviews appear, consumers read them and update their beliefs about products. Thus, in our model, substitution between products may occur when a new online review is published.

Figure 1 shows a simplistic example of how a review may influence consumer's beliefs about a given product feature, in this case *image quality*. The consumer has an initial belief distribution about the product quality, taking values from 0 to 1, with a mean value of 0.5. After reading a few reviews talking about the *excellent*, *fantastic*, and *superb* image quality, the belief distribution is updated and moves toward 1, having a mean value around 0.75. After reading a few more reviews, the belief is further updated, and so on.

Our paper touches a vast area of marketing research on conjoint analysis (Green and Srinivasan 1978) and preference measurement. The techniques presented in our paper and the statistical techniques used in conjoint analysis/preference measurement are targeted at determining how people value different features in a product or service. Where the approaches primarily differ is in the source of data used for analysis. Perhaps the simplest approach to preference elicitation, known as the self-explicated method, or SEM (Srinivasan 1988), is based on data from directly surveying consumers about their preferences for particular product attributes. Alternatively, one can use a simple conjoint analysis technique, in which a small set of attributes is used to create product profiles, and respondents are asked to directly rate these profiles. Because this approach does not scale well with the number of attributes, hybrid conjoint analysis techniques (Marshall and Bradlow 2002, Frenkel et al. 2002), the fast polyhedral method (Toubia et al. 2003), and the adaptive conjoint analysis (Johnson 1987) have been proposed in the literature.

A recent stream of research has focused on identifying new sources of data, such as revealed preference data, that can supplement stated preference data from the SEM and conjoint analysis. Prominent examples include combining scanner-based data with survey data (Horsky et al. 2006), incorporating market share information in choice-based conjoint analysis (Gilbride et al. 2008), and using Web product reviews to automate construction of conjoint attributes (Lee and Bradlow 2007). The paper by Lee and Bradlow (2007) is particularly close to our research. In their paper, the authors present an unsupervised text-mining technique to automatically identify attributes and attribute levels for conjoint analysis from the product review summaries posted on epinions.com. In contrast, in this paper, we present two alternative techniques for the attribute extraction: (i) a semisupervised extraction technique employing the crowdsourcing platform (Mturk) and (ii) a fully automated ontology-driven extraction technique. Potentially, both techniques can be applied in

Figure 1 Example of Sequential Belief Updating from Consumer Reviews



the same way as in Lee and Bradlow (2007), i.e., by using the extracted attributes and attribute levels as inputs to a conjoint analysis study. However, we take an alternative approach and show that reasonable information on user preferences can be extracted purely from contrasting temporal variation in aggregate sales data for a panel of consumer products with the emergence of new product reviews on the retailer's website.

To summarize, the main contribution of this paper is to show how textual information embedded in online reviews can be incorporated in a simple demand estimation model and to provide insights for using text-mining techniques in quantitative information systems, marketing, and economics research. Simultaneously, we aim to highlight the value of using an economic context to computer scientists to estimate both the intensity and the polarity of consumer opinions.

The rest of this paper is organized as follows. Section 2 presents our text-mining approach. Section 3 describes how we incorporate textual information in the empirical model of demand. We present a simple theoretical model of multiattribute choice under uncertainty that leads to our empirical estimation framework. A discussion of the results is given in §4. In §5, we show that the textual content can improve the power of purely predictive models of future sales and provide superior out-of-sample validation results. Finally, §6 concludes this paper with a discussion of the results, managerial implications, and directions for future research.

2. Econometric Modeling of Text Information

Prior research on consumer reviews and "word of mouth" marketing has largely ignored the nonquantitative nature of information contained in the consumer reviews. The economic analysis of textual data is nontrivial and presents a number of challenges. Consider, for example, our specific context: Consumers read product reviews and receive signals about different product attributes. To apply the model empirically, for each product review it is important to be able to answer the following three questions.

1. Which product features are evaluated in the product review?
2. What evaluations are given to these attributes?
3. What is the pragmatic and economic value of these evaluations to the consumer? That is, how are the evaluations taken into account by the consumer to adjust their beliefs about the given product?

In this section, we discuss the first two questions and our proposed text-mining solution. We present two alternative approaches for extracting the product features (§2.1) and opinions about these features (§2.2) from the text of product reviews: a fully automated approach, based on natural-language processing, and a crowdsourcing approach, using Amazon Mechanical Turk.

2.1. Identifying Important Product Features

The first step of our approach is to identify the product features that consumers describe in the reviews and determine which of them are important for the decision-making process of consumers. For the purpose of our study, it is not useful to follow product descriptions provided by manufacturers. This is because manufacturer-provided product descriptions are static and often do not contain information about intangible product features, such as the quality of product design, ease of use, robustness, and so on. Such intangible product features are hard to measure objectively, yet they may be important determinants of consumer buying decisions. Because we want to take consumer opinions explicitly into account, we do not exogenously specify the set of relevant product attributes. Instead, we rely on the contents of reviews to identify product features that are most *frequently discussed by consumers*.

2.1.1. Fully Automated Product Feature Identification Algorithm. Many techniques for identifying product features mentioned in consumer reviews

have been introduced in the last few years in text-mining research (Hu and Liu 2004, Ghani et al. 2006). One popular technique is to use a *part-of-speech* (*POS*) *tagger* to annotate each word in the review with its part of speech and mark whether the word is a noun, an adjective, a verb, and so on. Nouns and noun phrases are usual candidates for product features, although other constructs (like verb phrases) are used as well. Alternative techniques involve searching for statistical patterns in the text, for example, words and phrases that appear frequently in the reviews. Hybrid methods combine both approaches, where a POS tagger is used as a preprocessing step before applying an association-mining algorithm to discover frequent nouns and noun phrases.

Although it is generally acknowledged that the most frequently described features are nouns and noun phrases, in reality, reviewers do use a wide range of language constructs to describe the products. For example, consider the following sentence from a digital camera review: "A little noisy in low light, for example on cloudy days, grass will lack sharpness and end up looking like a big mass of green." This sentence gives an evaluation of the camera's picture quality even though the feature itself is never explicitly mentioned. Some techniques for discovering implicitly described product features have been developed. For example, one can use a binary classifier that determines whether a particular feature is discussed (implicitly) in the review or not (Ghani et al. 2006).

For our purposes, we follow the paradigm of (Hu and Liu 2004) and use a POS tagger to identify frequently mentioned nouns and noun phrases, which we consider to be candidate product features. Using WordNet (Fellbaum 1998), we then cluster these phrases into a set of similar nouns and noun phrases. In the final step, we examine the words that appear in a window of four words around the candidate noun phrase to extract the "context" in which a particular noun appears. Based on the context, we further group together the noun phrases that appear in similar contexts, using a hierarchical agglomerative clustering algorithm (Manning and Schütze 1999). The resulting set of clusters corresponds to the set of identified product features mentioned in the customer reviews.

Because of the inherent complexity of the natural language, no text-mining technique so far has proved to be as efficient in feature extraction as humans can be, especially when dealing with complex constructs such as implicitly described product features. Because the precision and recall of our text-mining technique can directly affect the quality of the results extracted by our econometric analysis (§3), it is important to consider alternative semiautomated feature extraction methods. We describe this below in the next subsection.

2.1.2. A Crowdsourcing-Based Technique for Product Feature Identification. To extract product features in a scalable, yet noise-free manner we decided to rely on a "human-powered computing" technique and used a semiautomated human intelligence approach instead of a fully automated approach. In particular, we used the *Amazon Mechanical Turk* system to distribute feature extraction assignments to workers. Amazon Mechanical Turk is an online marketplace, used to automate the execution of microtasks that require human intervention (i.e., cannot be fully automated using data-mining tools). Task requesters post simple microtasks known as human intelligence tasks (HITs) in the marketplace. Workers browse the posted microtasks and execute them for a small monetary compensation. The marketplace provides proper control over the task execution such as validation of the submitted answers or the ability to assign the same task to several different workers. It also ensures proper randomization of assignments of tasks to workers within a single task type.

The obvious question is whether such crowdsourcing techniques can be used for reliable extraction of information, given that it is difficult to check the quality of work submitted by each individual worker. The basic idea is to get each review examined by multiple workers and let the workers extract, *in free-text form*, the product features described in the review. If two workers extract the same product feature from the review, we consider the answer reliable. This idea has been used in the past, with a high degree of success, in the ESP game by von Ahn and Dabbish (2004). The goal in the ESP game is to get multiple users to tag images on the Web by letting them play a game: Two players, unknown to each other, see an image and have to type the same word to proceed to the next level. If they type the same word, they get points and proceed to the next image. The tagging results were of extremely high quality; the game is now licensed and used by Google (Google Image Labeler[1]). In the context of Mechanical Turk, Snow et al. (2008) review recent research efforts that use Mechanical Turk for annotation tasks and also evaluate the accuracy of "Turkers" for a variety of natural-language-processing tasks. They conclude that the nonexpert users of Mechanical Turk can generate results of comparable quality to those generated by experts, especially after gathering results for the same microtask using multiple Turkers. Sheng et al. (2008) describe how to effectively allocate tasks to multiple, noisy labelers (such as those on Mechanical Turk) to generate results that are comparable to those obtained with nonnoisy data.

[1] http://images.google.com/imagelabeler/.

Table 1 Product Features Identified in Each Product Category

Digital cameras	"Auto shake"/image stabilization, battery life, design, ease of use, flash, LCD, lens, megapixels, picture quality, shooting modes/variety of settings, size, video quality, zoom
Camcorders	Battery life, ease of use, LCD, picture/image quality, weight/size, video quality, audio quality, digital effects/enhancements, support of different output formats

In our work, we used similar principles and leveraged the workforce of Mechanical Turk for our task. To identify important product features in each of the three categories, we conducted a small pilot study. First, for each product category, we selected a random sample of 50 reviews. For each review we posted a HIT asking users to identify the product features described in the review and report them in free-text format; each review was processed by three independent workers. We paid each worker 50 cents per review, and the documents were processed within a couple of hours. The resulting list of the top 20 popular features for each product category in our data set is given in Table 1.

2.1.3. Empirical Comparison of Automated Text-Mining and Crowdsourcing-Based Approaches. We performed an additional pilot study to compare the performance (precision and recall) of the fully automated text-mining technique and the crowdsourcing-based technique with regard to the product feature extraction task. For the purpose of the pilot study, we used the top seven most popular features in the digital camera category and the top four most popular features in the camcorder category.[2] Furthermore, we randomly selected a set of 100 product reviews in each product category. Two human annotators carefully processed every review and every product feature to determine whether the feature was evaluated in a particular product review or not. We used the results of human annotators as the baseline for evaluating the feature extraction performance of both the fully automated and the crowdsourcing-based techniques. The corresponding precision and recall values are given in Tables 2 and 3. As we can see, both techniques demonstrated excellent predictive performance on the feature extraction task.

Table 2 Precision and Recall for the Digital Camera Data Set

Feature	Precision (automated)	Recall (automated)	Precision (crowdsourcing)	Recall (crowdsourcing)
Battery life	0.989	0.939	0.830	0.929
Design	0.760	0.974	0.816	0.782
Display	0.963	0.933	0.898	0.928
Ease of use	0.707	0.871	0.843	0.872
Picture quality	0.981	0.782	0.767	0.873
Size	0.741	0.927	0.787	0.894
Video quality	0.915	1.000	0.973	0.928

Table 3 Precision and Recall for the Camcorder Data Set

Feature	Precision (automated)	Recall (automated)	Precision (MTurk)	Recall (MTurk)
Ease of use	1.000	0.860	1.000	1.000
Picture quality	1.000	1.000	1.000	1.000
Size	0.832	0.911	0.950	0.890
Video quality	0.970	0.658	0.908	0.747

2.2. Identifying Customer Opinions

Of course, identifying product features per se is not the end goal. The important goal is to understand the customer's opinion about each of the identified product features. So, after identifying the product features, we need to identify users' opinions about the features that are embedded in the reviews. Each opinion is a phrase expressing reviewer's personal impression (usually based on prior experience) of the quality level of a certain product feature. Prior work has shown that in most cases consumers use adjectives, such as "*bad*," "*good*," and "*amazing*" to evaluate the quality of a product characteristic (Turney and Littman 2003, Hu and Liu 2004). The process of extracting user opinions can, in general, be automated. Following the automated approach, we can use a *syntactic dependency parser* to select the adjectives that refer to a noun or a phrase that we have identified as a product feature. An advantage of using a syntactic parser, as opposed to a single POS tagger, is that the syntactic parser can identify opinions that are "far" from the actual product feature.[3] This kind of an automated tool produces a set of noun phrases, for each review, that corresponds to pairs of product features and their respective evaluations contained in the review.

As in the case of extracting product features, in addition to the fully automated tool, we also consider a semiautomated crowdsourcing approach for the opinion phrase extraction. In the semiautomated approach, we used Amazon Mechanical Turk to extract the opinion phrases. We distributed reviews to the Mechanical Turk workers and asked two workers to process each review. Note that this is different from "standard coding" in that we do not have the same two workers labeling every simple piece of data.

[2] These are the same features that we later use in the discrete choice model.

[3] For example, in the phrase "the lens, which I bought from a website different than Amazon, is sharp and clear," the evaluations "sharp" and "clear" will be properly attributed to the feature "lens" by the syntactic parser, whereas the POS tagger will not capture such dependencies.

Instead we have hundreds of workers processing the data in parallel, and for quality assurance we require two workers to look at each piece. Each assignment contained the review text for a single product and a list of product features identified in the previous step. Workers were responsible for reading the review thoroughly and extracting opinion phrases evaluating any feature in the given list. The answers were returned in free-text format, and the workers were asked not to change the wording used in the original review. In our empirical study, interrater reliability was 34.27%, as measured by the Jaccard coefficient; that is, in more than one-third of all cases, two workers processing the same review reported *exactly the same* evaluation phrase for a particular product feature.[4]

3. Econometric Analysis

Our work is motivated by the seminal paper of Chevalier and Mayzlin (2006), who examined the dynamic effect of consumer product reviews on subsequent sales of books at Amazon.com and Barnesandnoble.com. Their estimation results show that the marginal effect of a one-star increase in the average review rating of a book on Amazon (as compared to the same book on Barnesandnoble.com) is equal to approximately 0.2 unit decrease in the logarithm of the sales rank. We build on their approach and proceed to evaluate how much consumer opinions about the different attributes of the product contribute to changes in product sales. Toward this, we examine "simple" hedonic products such as digital cameras and camcorders, which can be represented by a small number of well-defined attributes.

3.1. Data

We gathered data on a set of products using publicly available information at Amazon.com. The data set covered two different product categories: "digital cameras" (41 unique products) and "camcorders" (19 unique products). During a 15-month period (from March 2005 to May 2006), we collected daily price and sales rank information for the products in our data set using the programming interface provided by Amazon Web Services. Each observation contains the collection date, the product ID, the retail price on Amazon, the sales rank of the product, the product release date, and the average product rating according to the posted consumer reviews. Additionally, we used Amazon Web Services to collect the full set of reviews for each product. Each product review has a numerical rating on a scale of one to five stars, the date the review was posted, and the entire text posted by the reviewer.

Amazon.com does not publicly reveal information on actual product shares or total number of units sold for a particular product. Instead, Amazon reports a sales rank for each product, which can be used as a proxy for demand based on prior research (Brynjolfsson et al. 2003, Chevalier and Goolsbee 2003, Ghose and Sundararajan 2006). These studies have associated the sales ranks with demand levels for products such as books, 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 log-linear relationship $\ln(D) = a + b \cdot \ln(S)$, where D is the unobserved product demand, S is the observed sales rank, and $a > 0$, $b < 0$ are industry-specific parameters. However, for our purposes, such conversion is unnecessary; as long as one stays in the world of linear models, the estimation can be performed directly on sales ranks, and the marginal coefficients can be interpreted in terms of changes in sales ranks.

3.2. Empirical Model

In our data, we have a series of observations on sales and reviews for each product. Following Chevalier and Mayzlin (2006), we model the impact of product reviews on sales by directly incorporating product review information in a linear equation for the sales rank. Our estimation equation is given by

$$\log(s_{jt}) = d_j + \gamma_p p_{jt} + X_{jt}\beta^x_{jt} + Y_{jt}\beta^y_{jt} + Z_{jt}\beta^z_{jt} + \theta \log(s_{jt-1}) + \varepsilon_{jt}, \quad (1)$$

where s_{jt} is the sales rank for product j at time t, d_j is the product-specific fixed effect, p_{jt} is the price for product j at time t, X_{jt} is the vector of numeric review variables, Y_{jt} is the vector of textual review variables, and Z_{jt} is the vector of control variables. Note that the right side of Equation (1) includes only review content for products reviews that were published at least a day before the current time t; that is, instead of considering contemporaneous reviews, we consider a one-period lagged effect of reviews. The intuition behind this specification is that updating sales statistics on the Amazon's website takes some time, and hence the influence of "fresh" product reviews is

[4] Although this reliability score may be considered low for conventional surveys where participants report answers on numeric Likert-type scales, this is a good agreement score for free-text matching. Note that only for the "picture quality" feature did we identify 1,424 different evaluation phrases in consumer reviews, 197 of which were used more than once; it is significantly more difficult for two readers to select *exactly* the same phrase than to select the same number on a scale from one to five. Notice that we were checking for *identical* phrases to compute the inter-rater agreement and did not resort to substring or approximate matching. So this implies, for example, that "very good" and "very good!" will be considered nonidentical phrases.

unlikely to be captured by the current sales rank. In X_{jt} we include all standard numeric summaries for the set of product reviews: the average review rating, the total number of reviews, the total length of reviews, the fraction of one- and five-star reviews, and the standard deviation of review ratings to account for possible valence of reviews.

3.3. Theoretical Motivation for Empirical Estimation

Although Equation (1) is a direct extension of the approach of Chevalier and Mayzlin (2006), it also has an alternative independent construction based on a combination of two well-known theoretical approaches: multiattribute choice under uncertainty and Bayesian learning. We build such a model to formally motivate our empirical analyses. Although the theoretical model is not necessary to understand the methodologies and the results of this paper, through the description of the model, we hope to outline clearly the scope and applicability of our research, explain what the implicit assumptions behind our current approach are, and identify directions for future research. This can enable future researchers in this domain to adopt similar empirical approaches as ours. The full derivation of the model is given in the appendix; a short summary follows.

Products can be represented by n-dimensional tuples of attributes, and the quality of each attribute is uncertain to consumers. Consumers are expected-utility maximizers. We incorporate risk aversion by adopting negative exponential utility, a widely used specification (Roberts and Urban 1988, Bell and Raiffa 1988). To reduce uncertainty, consumers read product reviews before choosing a product and use Bayesian learning to update their beliefs about the quality of product attributes. Beliefs are assumed to be normally distributed, to be consistent with possibility of a recursive learning process (Chen 1985). It can be shown (Roberts and Urban 1988) that normal priors in combination with negative exponential utility give a particularly simple analytic representation for the expected utility function. The consumers' choices will be monotone with respect to the so-called "risk-adjusted preference function," which incorporates the linear component of the consumers' utility function evaluated at the mean of the current consumers' beliefs about the particular product quality and the additional risk-aversion component representing the cost of uncertainty about the product attributes. Our final result is obtained by connecting the "risk-adjusted preference function" directly to the market share for a particular product using Lemma 1.

In the scope of Equation (1), we can interpret vector Y_{jt} as representing the current mean of consumers' beliefs about the product quality. As new reviews are published, the change in Y_{jt} represents the shift in consumers' beliefs, whereas the change in $Y_{jt}\beta^y_{jt}$ represents the corresponding direct effect of these changes on the product sales. The change in the risk-aversion component is controlled for by including additional variables in the regression such as the fraction of one- and five-star reviews and the standard deviation of review ratings.

3.4. Incorporating Textual Information

Every component in vector Y_{jt} represents a single possible opinion phrase, i.e., a combination of an evaluation e (for example, "good," "bad," "excellent") and a product feature f ("zoom," "size," "weight"). In the following discussion, we use $\mathcal{F}$ to represent the set of all interesting product features and $\mathcal{E}$ to represent the set of all interesting evaluations. Then the dimension of vector Y_{jt} will be equal to $\|\mathcal{F}\| \times \|\mathcal{E}\|$. We use $Y_{jt}(f, e)$ to represent a component corresponding to the pair of feature f and evaluation e, and $Score(f, e)$ to represent the corresponding slope in β^y_{zt} (the interpretation is that this value is a "score" that consumers assign to this particular opinion phrase). We can now write Equation (1) as

$$\log(s_{jt}) = d_j + \gamma_p p_{jt} + X_{jt}\beta^x_{jt} + \sum_{f\in\mathcal{F}}\sum_{e\in\mathcal{E}} Y_{jt}(f, e)Score(f, e) + Z_{jt}\beta^z_{jt} + \theta\log(s_{jt-1}) + \varepsilon_{jt}. \quad (2)$$

Equation (2) has an interesting and novel interpretation. Note that traditional consumer-review-mining approaches consider extracted product features and opinions as simple sets and impose no algebraic structure on them. We propose that we can *meaningfully* define a vector space structure for consumer reviews. Each opinion phrase (for example, "great synchronization with PC") will represent a single dimension of a consumer review. Furthermore, we propose measuring the value of each dimension as the number of times the corresponding opinion phrase occurred in the review text, normalized by the number of times the corresponding feature was evaluated in the review text. A theoretical justification for such weighting scheme based on a simple model of Bayesian learning by consumers with normal priors can be found in the appendix. The proposed weighting scheme is

$$Y_{jt}(f, e) = \frac{N(f, e)}{s + \sum_{\hat{e}\in\mathcal{E}} N(f, \hat{e})}. \quad (3)$$

This idea can be illustrated with a simple example.

Example 1. Consider the following review for a digital camera: "The camera is of high quality and relatively easy to use. The lens is fantastic. Bright and clear! I have been able to use the LCD

viewfinder.... To summarize, this is a very high quality product." This review can be represented by elements of the consumer review space with the following weights (assume $s = 0$ for this example): $Y_{jt}(quality, high) = 1/1 = 1.0$, $Y_{jt}(use, easy) = 1/1 = 1.0$, $Y_{jt}(lens, fantastic) = 1/3 = 0.33$, $Y_{jt}(lens, bright) = 1/3 = 0.333$, $Y_{jt}(lens, clear) = 1/3 = 0.333$. Notice that each opinion phrase dimension has a weight coefficient determining its relative importance in the review. Because the feature *quality* is evaluated once in the review (*"high quality"*), the weight of the evaluation is 1.0. In contrast, the feature *lens* has three evaluations (*fantastic*, *bright*, *clear*); therefore the weight of each evaluation is 1/3.

If we employ this particular representation of consumer reviews, the impact of the product reviews on the market share of that product can be modeled simply as a linear functional from the space of consumer reviews.

3.5. Identification

The typical source of endogeneity in demand models is unobservable exogenous shocks that simultaneously affect prices set by firms as well as buying decisions made by consumers. In addition, there could be some external factors that influence both consumer reviews and product demand, such as advertising or publicity. Thus, using ordinary least squares estimation, we will likely overestimate the direct effect of consumer reviews on product demand. To alleviate this concern, we use data on the "*product search volume*" of different products from Google Trends to control for exogenous demand shocks. For each product, we retrieved the search volume from the Google Trends website. Because the search volume for the brand can be correlated with the product sales, we include it as a control variable in the model. The use of search volume from Google Trends as a measure of product publicity acts as suitable control for any unobserved factor driving both sales and word of mouth; it is consistent with the approach of Luan and Neslin (2009), who show that publicity has a significant impact when mapping the relationship between sales and word of mouth. Additionally, we follow Villas-Boas and Winer (1999) and use lagged product prices as instruments. The lagged price may not be an ideal instrument because its possible to have common demand shocks that are correlated over time and affect prices set by producers. Nevertheless, common demand shocks that are correlated through time are essentially trends. Our control for trends using Google search volume data thus should alleviate most, if not all, such concerns.

Furthermore, our data set represents a longitudinal panel in which a number of products have been observed for more than a year. Hence, we need to control for different time-series-specific effects such as autocorrelation in the sales rank. Toward this, we include a lag of the dependent variable in the model and apply the system GMM (Hansen 1982) estimator for dynamic panel data models developed by Arellano and Bover (1995). The system GMM estimator uses the original estimation equation to obtain a system of two equations: one in differences and one in levels. The system GMM estimator has been shown by Blundell and Bond (1998) to have much better finite-sample properties than that of the original difference GMM estimator.[5] We apply the finite-sample correction proposed by Windmeijer (2005), which corrects for the two-step covariance matrix and increases the efficiency of the GMM estimator. We were careful to take into account the problem of using too many lags as instruments (Roodman 2006).

3.6. Dealing with "Curse of Dimensionality"

Because the set of different opinion phrases extracted from online user-generated content is typically very large, it is infeasible to include all these variables in any statistical model. One has to restrict consideration to only the top K most popular opinion phrases in each product category, for some relatively small K. Unfortunately, this also means that, after the model estimation, we might get coefficients that also *include the projection of some omitted variables*. That is particularly problematic, because many of the omitted variables are negative, whereas many frequent phrases are generally positive (see Table 4). This happens because consumers tend to use standard opinion phrases to describe their positive impressions but use longer and comparatively far less standardized sentences to describe their negative experiences. For example, for digital camera products, highly positive evaluations of the "picture quality" feature might frequently co-occur with negative evaluations of the camera size in consumer reviews (like a person saying "this camera has great picture quality but its too big and heavy"). Including the "great picture quality" opinion but excluding the "big and heavy" opinion from the model will likely bias downward our estimates of consumer value for "great picture quality."

[5] Arellano and Bond (1991) developed a GMM estimator that treats the model as a system of equations, one for each time period. The equations differ only in their instrument/moment condition sets. The key idea is that if the error terms are serially uncorrelated, then the lagged values of the dependent variable and the exogeneous variables represent valid instruments. The resulting estimator is known as the difference GMM (DGMM). A potential difficulty with the DGMM estimator is that lagged levels may not be good instruments for first differences when the underlying variables are highly persistent over time.

Table 4 Top 20 Most Frequent Product Opinions Identified in "Digital Camera" and "Camcorder" Product Categories

Camera feature	Evaluation	Freq.	Camcorder feature	Evaluation	Freq.
Ease of use	Easy	405	User friendliness	Easy	112
Picture quality	Great	354	Size/weight	Small	72
Size	Small	321	Video quality	Great	48
Ease of use	Easy to use	160	Video quality	Excellent	43
Picture quality	Good	151	Size/weight	Compact	38
Picture quality	Excellent	120	Video quality	Good	38
Size	Compact	115	Picture/image quality	Good	36
Picture quality	Clear	92	Picture/image quality	Excellent	35
LCD	Large	85	User friendliness	Easy to use	30
Size	Light	78	User friendliness	Great	23
Picture quality	Sharp	71	Different output formats	Minidv	22
Picture quality	Blurry	70	Audio quality	Good	21
Design	Great	65	Size/weight	Light	20
Ease of use	Very easy	65	Size/weight	Lightweight	19
Picture quality	Amazing	63	Picture/image quality	Nice	18
Video quality	Great	59	Picture/image quality	Clear	17
Ease of use	Simple	57	User friendliness	Very easy	17
Size	Little	54	Picture/image quality	Very good	17
Picture quality	Crisp	54	Video quality	Very good	17
Battery life	Good	48	Picture/image quality	Great	15

3.6.1. Solution 1: Clustering Opinion Phrases. The first solution we propose for the problem of omitted variables is based on a simple idea of learning and exploiting similarity between different opinion phrases. We propose a nonparametric, data-centric approach to keep the number of regressors small. The idea is to retain the top K ($K = 20$ in our application) most popular opinion phrases for each product category and then perform clustering or remapping of omitted opinions. In our case, we use a technique based on statistical properties of the data.[6] In particular, we use the concept of *pointwise mutual information* (*PMI*) (Turney 2002) to measure the distance between two opinions. For each evaluation, such as "out of focus" applied to "picture quality," we calculated its PMI value with all top 20 regressors using the following formula:

$$PMI(f, e_1, e_2) = \frac{Count(f, e_1, e_2)}{Count(f, e_1)Count(f, e_2) + s}, \quad (4)$$

where $Count(f, e_i)$ is the number of consumer reviews containing the evaluation e_i for feature f, $Count(f, e_i, e_j)$ is the number of consumer reviews containing both evaluation e_i and e_j for feature f, and s is some smoothing constant. Finally, we mapped all the evaluations in the tails to their nearest neighbors using the PMI distance method. In our example, "out of focus" picture quality will be mapped to "blurry" picture quality as shown in the Table 7. The table lists a subsample of mappings used for "digital cameras."

3.6.2. Solution 2: Using Externally Imposed Polarity for Evaluation. The second solution is to exogenously assign explicit polarity semantics to each evaluation word, for example, to consider "excellent" to have a value of 0.9, "bad" to be −0.5, "horrible" to be −0.9, and so on. This solution effectively reduces the number of coefficients to evaluate to the number of different product features by exploiting additional domain knowledge. To implement it, we extracted all evaluations that were associated with the product features that we considered for each category. Then, we used Amazon Mechanical Turk to create our ontology, with the scores for each evaluation phrase. Our process for creating these "external" scores was done as follows. We asked nine Mechanical Turk workers to look at the pair of the evaluation phrase together with the product feature and assign a grade from −3 (strongly negative) to +3 (strongly positive) to the evaluation. This resulted in a set of nine independently submitted evaluation scores; we dropped the highest and lowest evaluation scores, and used the average of the remaining seven evaluations as the externally imposed score for the corresponding evaluation–product phrase pair. We should stress that the scoring of the evaluation phrases only needs to be done once per product category, because the set of product features and the corresponding evaluation phrases are highly unlikely to change over time.

In §4.3, we discuss some pros and cons of using the imposed polarity approach vs. the inferred polarity approach.

[6] Of course, other clustering techniques may be used, perhaps exploiting resources such as WordNet or search engines.

Table 5 Descriptive statistics for Digital Cameras

Variable	Obs.	Mean	Std. dev.	Min	Max
log(*sales rank*)	11,897	5.38	1.77	0	10.79
log(*price*)	11,897	5.70	0.657	3.434	7.3139
Rating	11,897	3.26	1.798	1	5
Trend	11,897	4.26	0.281	2.89	4.846
log(*numreviews*)	11,897	3.37	1.09	0	4.93
Product age	11,897	204.31	114.09	1	423
LCD, large	11,897	3.662	4.1562	0	19
LCD, large^2	11,897	30.678	53.416	0	361
Battery life, good	11,897	1.094	1.836	0	10
Battery life, good^2	11,897	4.566	13.1679	0	100
Design, great	11,897	1.181	1.67	0	7
Design, great^2	11,897	4.188	7.61	0	49
Ease of use, easy	11,897	7.576	7.819	0	36
Ease of use, easy to use	11,897	2.852	3.342	0	20
Ease of use, easy to use^2	11,897	19.297	43.615	0	400
Ease of use, easy^2	11,897	118.503	199.276	0	1,296
Ease of use, simple	11,897	1.177	1.743	0	10
Ease of use, simple^2	11,897	4.423	11.603	0	100
Ease of use, very easy	11,897	2.012	3.2546	0	20
Ease of use, very easy^2	11,897	14.635	47.8091	0	400
Picture quality, amazing	11,897	2.073	2.6145	0	10
Picture quality, amazing^2	11,897	11.129	19.7314	0	100
Picture quality, blurry	11,897	2.093	3.2692	0	18
Picture quality, blurry^2	11,897	15.06525	41.4147	0	324
Picture quality, clear	11,897	5.173	6.019	0	23
Picture quality, clear^2	11,897	62.977	112.311	0	529
Picture quality, crisp	11,897	2.341	3.212	0	12
Picture quality, crisp^2	11,897	15.798	31.604	0	144
Picture quality, excellent	11,897	3.15	3.32	0	15
Picture quality, excellent^2	11,897	21.021	34.075	0	225
Picture quality, good	11,897	2.759	3.647	0	19
Picture quality, good^2	11,897	20.91	53.001	0	361
Picture quality, great	11,897	8.15	9.426	0	39
Picture quality, great^2	11,897	155.364	284.025	0	1,521
Picture quality, sharp	11,897	2.899	2.8411	0	14
Picture quality, sharp^2	11,897	16.472	24.655	0	196
Picture quality, very good	11,897	1.454	2.68	0	10
Picture quality, very good^2	11,897	9.29	24.299	0	100
Size, compact	11,897	5.41	5.668	0	23
Size, compact^2	11,897	61.40	99.727	0	529
Size, light	11,897	3.225	4.22	0	16
Size, light^2	11,897	28.203	54.221	0	256
Size, small	11,897	7.84	8.28	0	33
Size, small^2	11,897	130.0022	182.54	0	1,089
Video quality, great	11,897	1.47	2.5345	0	10
Video quality, great^2	11,897	8.58	19.60	0	100

4. Results

In this section, we discuss the estimation results for each product category. We start by discussing results from the model with inferred polarity and then proceed to discuss results from the model with exogenously imposed polarity.

4.1. Inferred Polarity Model

Because of the limited size of our sample, we used only the top 20 most popular opinion phrases for the "digital camera" category and the top 10 most popular opinion phrases for the "camcorder" category. The rest of opinion phrases were mapped into one of the top opinions based on the clustering algorithm described in §3.6. Results are given in Tables 9, 10, and 11.

The first column of each of these tables reports results of a GMM estimation without textual data.[7] Following Villas-Boas and Winer (1999), we used lagged prices to instrument for potential price endogeneity. The second column reports estimates from the same model but including textual data. The third

[7] We have also estimated a simple IV (two-stage least squares) model with similar set of instruments. Because the results are qualitatively very similar to our current results, we only present results from the more efficient GMM estimator.

Table 6 Descriptive Statistics for Camcorders

Variable	Obs.	Mean	Std. dev.	Min	Max
log(*sales rank*)	6,786	5.94	1.25	1.61	9.47
log(*price*)	6,786	5.94	0.88	3.044	7.004
Rating	6,786	2.543	1.868	1	5
Trend	6,786	4.454	0.171	3.618	4.859
log(*numreviews*)	6,786	2.252	0.976	0	4.331
Product age	6,786	195.85	115.557	1	420
Picture/image quality, excellent	6,786	1.066	1.573	0	6
Picture/image quality, excellent^2	6,786	3.61	7.97	0	36
Picture/image quality, good	6,786	1.846	2.829	0	11
Picture/image quality, good^2	6,786	11.405	29.515	0	121
Picture/image quality, great	6,786	1.510	2.464	0	14
Picture/image quality, great^2	6,786	8.347	26.4101	0	196
Size/weight, compact	6,786	2.142	3.2608	0	12
Size/weight, compact^2	6,786	15.211	35.87	0	144
Size/weight, small	6,786	3.515	3.91	0	16
Size/weight, small^2	6,786	27.628	47.889	0	256
User friendliness, ease of use, easy	6,786	4.561	6.068	0	22
User friendliness, ease of use, easy to use	6,786	1.090802	1.2737	0	4
User friendliness, ease of use, easy to use^2	6,786	2.81	4.346	0	16
User friendliness, ease of use, easy^2	6,786	57.592	121.84	0	484
Video quality, excellent	6,786	1.35	2.071	0	9
Video quality, excellent^2	6,786	6.11	16.77	0	81
Video quality, good	6,786	1.784	2.552	0	9
Video quality, good^2	6,786	9.689	21.398	0	81
Video quality, great	6,786	1.738	2.251	0	12
Video quality, great^2	6,786	8.084	21.131	0	144

column adds a lag of the dependent variable to control for autocorrelation in the sales rank and applies the Arellano and Bover (1995) estimator. The last column consists of estimates from a robustness check in which we aggregate observations on a weekly basis to ensure that using daily units does not result in a significant downward bias of the standard errors because of potential within-group autocorrelation in daily regression residuals (Moulton 1986).

GMM estimators for dynamic panel data models such as the Arellano and Bover (1995) estimator with default settings will use, for each time period, all available lags of the specified variables as instruments, thus generating moment conditions prolifically (Roodman 2006). To avoid model overidentification, we have restricted the number of lags to two. The Sargan (1958) test of the moment conditions does not indicate overidentification ("digital camera" data set,

Table 7 Some Mappings Produced by PMI Based Clustering in the "Digital Cameras" Category

From feature	From evaluation	To feature	To evaluation	PMI (∗Const)
Picture quality	Dark	Picture quality	Blurry	43.25259516
Ease of use	Very easy to use	Ease of use	Very easy	41.2371134
Battery life	Very good	Battery life	Good	31.91489362
Picture quality	Very clear	Picture quality	Clear	30.6122449
Picture quality	Vivid	Picture quality	Sharp	29.12621359
Picture quality	Grainy	Picture quality	Blurry	25.95155709
Picture quality	Vibrant	Picture quality	Crisp	22.98850575
Picture quality	Bright	Picture quality	Clear	21.78649237
Picture quality	Fuzzy	Picture quality	Blurry	18.33740831
Picture quality	Detailed	Picture quality	Clear	18.11594203
LCD	Bright	LCD	Large	17.94974073
Video quality	Fantastic	Video quality	Great	16.39344262
Picture quality	Fabulous	Picture quality	Clear	16.33986928
Size	Good	Size	Light	16.28664495
Picture quality	Out of focus	Picture quality	Blurry	15.97444089
Design	Perfect	Design	Great	15.97444089
Picture quality	Blurred	Picture quality	blurry	15.59251559
Ease of use	Great	Ease of use	Simple	15.41623844

Table 8 Predictive Accuracy and Area Under the ROC Curve for the Sales Rank Classifier

Model	Group	AUC
No text	Digital camera	0.574
Text	Digital camera	0.644
No text	Camcorder	0.544
Text	Camcorder	0.617

Notes. The dependent variable is +1 or −1 if the sales rank goes up/down within the next week. The reported numbers are averages from the 10-fold cross-validation. The baseline AUC is a random score of 0.5.

$\chi^2(174) = 167.4328$, $p = 0.6258$; "camcorder" data set, $\chi^2(177) = 188.0764$, $p = 0.2702$).

Our first key research objective is to investigate whether textual information embedded in product reviews influences purchases beyond the use of numeric ratings. For each data set, we conducted a Wald test of joint significance for the coefficients on the textual variables using the estimates of the GMM model. The test rejects the null hypothesis at the 1% significance level ("digital camera" data set, $\chi^2(20) = 78.44$, $p = 0.000$; "camcorder" data set, $\chi^2(10) = 50.03$, $p = 0.000$).

We can make several inferences from the regression coefficients. Note that the signs of the coefficients of the numeric variables are in accordance with what one would expect. The coefficient on price is positive and significant, implying that higher product prices increase the sales rank and therefore decrease product sales. The coefficient on age is also positive, implying that products sales tend to decrease with time. Consistent with Chevalier and Mayzlin (2006), we find a positive effect of the average review rating on the product sales in both categories. Note that in all four cases, the absolute value of the coefficient on the average review rating goes down when the textual data are incorporated in the model. For instance, for "digital camera" data set, the no-text model reports a coefficient of −1.04 for the average review rating, whereas the next three models incorporating text data show significantly smaller magnitude of the effect of the average rating (in the neighborhood of −0.2). We interpret this result as a partial evidence in favor of the hypothesis that consumers' shopping decisions are not only affected by the average product rating, but by the actual textual contents of the reviews.

The volume of reviews shows a positive effect on product sales in both categories. This is consistent with classical models of risk aversion: Given two similar products with similar average review ratings,

Table 9 GMM and Dynamic GMM Estimation Results for Digital Camera Category with Inferred Polarity, Part A

	(1) Model 1 (GMM)	(2) Model 2 (GMM)	(3) Model 3 (DGMM)	(4) Model 4 (DGMM)
Price (unit = \$100)	0.101***	0.0954***	0.0517***	0.0155
	(9.26)	(8.87)	(4.13)	(0.94)
Trend	−1.440***	−0.858***	−0.291***	−0.419***
	(−63.56)	(−32.43)	(−7.59)	(−4.38)
Age	0.00954***	0.00855***	0.00330***	0.00460***
	(135.75)	(123.57)	(15.53)	(8.18)
Isused	0.628***	0.560***	0.216***	0.340***
	(65.33)	(52.91)	(11.87)	(7.53)
Fraction of one-star reviews	−0.571***	−0.0234	0.103	0.234
	(−5.83)	(−0.23)	(0.80)	(0.89)
Fraction of five-star reviews	2.340***	0.175*	0.0146	0.405*
	(60.28)	(2.54)	(0.17)	(2.22)
Numreviews (unit = 10)	−0.358***	−0.213***	−0.0716***	−0.117***
	(−59.05)	(−33.87)	(−7.86)	(−4.98)
Rating	−1.040***	−0.279***	−0.0820*	−0.223*
	(−43.49)	(−9.93)	(−2.25)	(−2.57)
Reviewlength (unit = 10,000 words)	0.0397***	−0.0976***	−0.0486***	−0.0448***
	(5.69)	(−14.96)	(−6.16)	(−3.30)
Ratingstdev	−0.722***	−0.311***	−0.0947***	−0.240***
	(−46.46)	(−16.40)	(−3.81)	(−4.02)
Hasrating	2.542***	0.615***	0.206	0.392
	(30.97)	(7.01)	(1.80)	(1.44)
		…	…	…
		…	…	…

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Table 10 GMM and Dynamic GMM Estimation Results for Digital Camera Category with Inferred Polarity, Part B

	(1) Model 1 (GMM)	(2) Model 2 (GMM)	(3) Model 3 (DGMM)	(4) Model 4 (DGMM)
Picturequalityblurry		−0.524***	−0.214**	−0.187
		(−8.95)	(−2.73)	(−1.11)
Picturequalityclear		−0.819***	−0.288**	−0.517**
		(−11.66)	(−3.17)	(−2.66)
Picturequalitygood		0.507***	0.232***	0.352**
		(12.67)	(4.41)	(2.90)
Picturequalityverygood		−0.654***	−0.366**	−0.863**
		(−7.73)	(−3.09)	(−2.84)
Picturequalitygreat		−3.043***	−1.186***	−1.634***
		(−68.58)	(−12.93)	(−6.62)
Picturequalitysharp		0.486***	0.169	0.535*
		(6.09)	(1.65)	(2.25)
Picturequalitycrisp		0.0846	0.0841	−0.375
		(0.87)	(0.69)	(−1.41)
Picturequalityexcellent		−1.755***	−0.692***	−0.934***
		(−23.40)	(−6.80)	(−4.26)
Picturequalityamazing		−4.901***	−1.921***	−2.194***
		(−39.32)	(−10.36)	(−5.79)
Easeofusesimple		−3.086***	−1.349***	−1.330***
		(−33.39)	(−9.57)	(−4.97)
Easeofuseeasy		−0.540***	−0.221***	−0.374**
		(−12.16)	(−3.80)	(−2.89)
Easeofuseeasytouse		1.706***	0.710***	0.550*
		(19.27)	(5.96)	(2.20)
Easeofuseveryeasy		2.208***	0.824***	1.470***
		(18.41)	(5.18)	(3.88)
Sizesmall		−0.699***	−0.237***	−0.321***
		(−24.28)	(−5.92)	(−3.41)
Sizecompact		−0.530***	−0.197**	−0.412**
		(−10.26)	(−3.03)	(−3.06)
Sizelight		0.0160	−0.00451	−0.155
		(0.24)	(−0.05)	(−0.84)
Videoqualitygreat		3.102***	1.196***	1.634***
		(28.40)	(7.60)	(4.53)
LCDlarge		−0.326***	−0.139***	−0.156*
		(−13.07)	(−4.18)	(−2.13)
Designgreat		−2.401***	−0.872***	−1.005**
		(−21.55)	(−5.79)	(−2.82)
Batterylifegood		5.624***	2.224***	2.744***
		(44.76)	(10.90)	(6.14)
Log of *sales rank*			0.618***	0.466***
			(26.32)	(7.23)
_Cons	5.242***	5.908***	2.123***	3.433***
	(62.26)	(67.76)	(11.96)	(7.29)
N	7,307	7,307	7,267	1,349

Note. The *t*-statistics are in parentheses.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

consumers will prefer the product that was reviewed more. Controlling for the number of reviews, we cannot make a similar claim for the total review length. Whereas in the "digital camera" category the coefficient on the review length is negative (indicating a positive effect on sales), in the "camcorder" category, it is not statistically significant. This ambiguous result may be due to two conflicting phenomena: Although longer reviews may theoretically provide more information about a product, they may also be perceived as more bloated and less relevant or helpful. Thus, everything else being equal, consumers may have a

Table 11 GMM and Dynamic GMM Estimation Results for Camcorder Category with Inferred Polarity

	(1) Model 1 (GMM)	(2) Model 2 (GMM)	(3) Model 3 (DGMM)	(4) Model 4 (DGMM)
Price (unit = \$100)	0.0560***	0.0374*	0.0272	0.0341
	(3.55)	(2.30)	(1.70)	(1.44)
Trend	0.687***	0.605***	0.326***	0.438***
	(17.28)	(13.87)	(6.59)	(3.80)
Age	0.00688***	0.00580***	0.00342***	0.00283***
	(85.05)	(64.42)	(19.28)	(7.51)
Isused	0.341***	0.250***	0.141***	0.152**
	(17.42)	(11.39)	(5.84)	(3.13)
Fraction of one-star reviews	−3.362***	−2.886***	−1.633***	−1.252***
	(−28.84)	(−23.20)	(−10.84)	(−4.26)
Fraction of five-star reviews	1.746***	0.867***	0.543***	0.433**
	(21.61)	(16.13)	(7.96)	(3.15)
Numreviews (unit = 10)	−0.461***	−0.228***	−0.134***	−0.103
	(−32.51)	(−8.93)	(−4.67)	(−1.85)
Rating	−1.017***	−0.736***	−0.446***	−0.417***
	(−47.93)	(−28.52)	(−12.45)	(−5.88)
Reviewlength (unit = 10,000 words)	0.122***	0.0000268	0.00278	−0.00547
	(8.15)	(0.00)	(0.14)	(−0.13)
Ratingstdev	−1.024***	−0.499***	−0.292***	−0.219***
	(−46.77)	(−24.92)	(−10.73)	(−3.83)
Hasrating	1.536***	1.890***	1.158***	1.201***
	(20.80)	(24.10)	(11.23)	(5.50)
Pictureimagequalitygood		−2.523***	−1.444***	−1.125***
		(−36.73)	(−13.91)	(−5.12)
Pictureimagequalitygreat		0.245	0.236	0.248
		(1.28)	(1.16)	(0.72)
Pictureimagequalityexcellent		−1.180***	−0.791***	−0.548
		(−5.81)	(−3.39)	(−1.37)
Sizeweightsmall		0.117	0.0544	0.0147
		(1.54)	(0.66)	(0.09)
Sizeweightcompact		0.115	0.134	0.0743
		(1.92)	(1.89)	(0.46)
Userfriendlinesseaseofuseeasy		−1.201***	−0.701***	−0.539***
		(−14.78)	(−7.51)	(−3.35)
Userfriendlinesseaseofuseeasytouse		−1.816***	−1.074***	−0.610***
		(−29.81)	(−12.57)	(−3.51)
Videoqualitygood		0.383***	0.241*	0.221
		(3.50)	(2.11)	(1.02)
Videoqualitygreat		1.389***	0.771***	0.587**
		(14.00)	(7.15)	(2.89)
Videoqualityexcellent		−0.426***	−0.186*	−0.142
		(−5.85)	(−2.19)	(−0.87)
Log of *sales rank*			0.424***	0.551***
			(16.43)	(9.33)
_Cons	7.309***	6.257***	3.555***	2.587***
	(60.49)	(57.49)	(17.45)	(5.99)
N	4,377	4,377	4,356	790

Note. The *t*-statistics are in parentheses.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

preference for shorter, more readable reviews, which is consistent with Ghose and Ipeirotis (2010).

Finally, the standard deviation of the set of review ratings has a strong positive effect on sales in both categories. This finding suggests that controlling for the average review rating, consumers will prefer a more polarized set of reviews. For example, a set of a one-star and a five-star review will be preferred to a set of two three-star reviews. We argue that a more polarized set of reviews may be perceived as more informative

by consumers, consistent with prior work (Ghose and Ipeirotis 2010).

The more interesting results, however, are related to the coefficients for the text-based data. In the "digital camera" category, the top positive evaluations are (in decreasing order of importance) "amazing picture quality," "great picture quality," "simple ease of use," "great design," "excellent picture quality," "very good picture quality." In the "camcorder" category, the statistically significant positive evaluations are (in decreasing order of importance) "good picture/image quality" and "ease of use." Interestingly, some seemingly positive evaluations are estimated to be negative and statistically significant: "good" and "sharp picture quality" and "good battery life" for digital cameras, and "great video quality" for camcorders. One plausible explanation for this effect is that consumers of camcorders have *strong prior expectations* about the product video quality, just as consumers of digital cameras have *strong prior expectations* about the battery life and the picture quality, and it is the *difference between the signal and the prior expectation that determines the influence of a consumer review* (see the appendix). Therefore, it is possible that this may be due to the fact that, from the buyer's perspective, a neutral or a lukewarm evaluation for a major product attribute is not sufficiently strong enough to warrant an increase in sales.

4.2. Exogenously Imposed Polarity Model

We also estimated the model in which the polarity of evaluations is imposed ex ante from the predefined ontology. In this scenario, the data are used only to infer the weights of the product features. The results are given in Tables 12 and 13. Consistent with the inferred polarity model, the textual data are jointly significant at the 1% level ("digital camera" data set, $\chi^2(7) = 62.51$, $p = 0.000$; "camcorder" data set, $\chi^2(4) = 35.05$, $p = 0.000$). Furthermore, the Sargan (1958) test alleviates any concerns of overidentification ("digital camera" data set, $\chi^2(174) = 165.6693$, $p = 0.6620$; "camcorder" data set, $\chi^2(174) = 190.2418$, $p = 0.2351$).

Results for the control variables align very well with expectations and are consistent with the results from the induced polarity model. Both price and age have a negative effect on product sales, whereas the average review rating, the volume of reviews, and the standard deviation of review ratings have a positive impact on sales.

Finally, the signs on textual variables are mostly as expected, with a couple of interesting exceptions. In the "digital camera" category, "picture quality" is not recognized as a statistically significant feature, and "ease of use" seems to have a negative effect on sales. In the "camcorder" category, "video quality" has a similar problem. As before, we argue that the effect may be purely due to strong prior consumer expectations for certain features of certain products. For example, if consumers expect digital cameras to have good picture quality by default, then its effect on sales is unlikely to be statistically significant. Nonetheless, it can also be an indication of other limitations in the data or estimation approach. We discuss these below.

4.3. Interpretation of Results: Limitations

We believe it is useful to alert the readers to potential limitations of our approach and how such limitations can affect interpretation of the results. First of all, heterogeneity in consumer tastes and its interaction with producers' pricing strategies can potentially bias the estimates of a simple linear model. Pakes (2003) shows that if firms are engaging in Bertrand pricing, then markups of products over marginal costs are a complex function of the interplay between characteristics of competing products and the distribution of consumer preferences. As a result, the coefficients on product characteristics obtained by linear estimators may often be in an unexpected direction (i.e., increases in good features may be negatively correlated with price).[8]

Second, omitted variable bias can also be present in our approach. Extracting all opinions in reviews with complete precision using automated methods is practically implausible (as is widely recognized in the text-mining community). Even if one resorts to manual processing of reviews, there are distinct trade-offs: Because of the limited number of observation points in manual processing, one either has to drop many extracted opinions to avoid the curse of dimensionality (thus creating an omitted variable bias), or one has to pool multiple opinions together (thus biasing the results in a different way).

For instance, in the situation with "good battery life," an unexpected sign of the estimate can likely be attributed to a mix of high prior expectations about the product quality combined with the omitted variable bias. We investigated this further. A manual inspection of the review corpora shows that in many cases when consumers use the "good battery life" opinion, they often proceed with a complex critique of the camera using opinion sentences or describing features that cannot be fully captured in the model. Some examples of such scenarios using actual phrases extracted from review data are given below (original spelling preserved, italics added):

- "The battery life *even though did not impress me* was still good and presence of battery level indicator instead of low battery light was also a big plus."
- "Battery life has been good, but *not great*."

[8] We thank an anonymous reviewer for suggesting this explanation.

Table 12 GMM and Dynamic GMM Estimation Results for Digital Camera Category with Induced Polarity

	(1) Model 1 (GMM)	(2) Model 2 (GMM)	(3) Model 3 (DGMM)	(4) Model 4 (DGMM)
Price (unit = \$100)	0.101***	0.101***	0.0513***	0.0186
	(9.26)	(9.10)	(4.07)	(1.12)
Trend	−1.440***	−0.941***	−0.341***	−0.410***
	(−63.56)	(−38.04)	(−9.07)	(−4.52)
Age	0.00954***	0.00889***	0.00350***	0.00482***
	(135.75)	(139.20)	(15.84)	(8.30)
Isused	0.628***	0.674***	0.263***	0.383***
	(65.33)	(66.46)	(13.14)	(7.86)
Fraction of one-star reviews	−0.571***	−0.718***	−0.160	−0.0174
	(−5.83)	(−6.94)	(−1.26)	(−0.06)
Fraction of five-star reviews	2.340***	1.287***	0.490***	0.920***
	(60.28)	(23.49)	(6.46)	(5.00)
Numreviews (unit = 10)	−0.358***	−0.355***	−0.130***	−0.200***
	(−59.05)	(−50.27)	(−10.77)	(−6.36)
Rating	−1.040***	−0.283***	−0.0934**	−0.248**
	(−43.49)	(−10.43)	(−2.72)	(−2.95)
Reviewlength (unit = 10,000 words)	0.0397***	−0.00686	−0.0117	0.00254
	(5.69)	(−1.04)	(−1.55)	(0.19)
Ratingstdev	−0.722***	−0.193***	−0.0576*	−0.207***
	(−46.46)	(−10.61)	(−2.53)	(−3.70)
Hasrating	2.542***	1.072***	0.399***	0.614*
	(30.97)	(12.35)	(3.57)	(2.27)
Picturequality		0.0303	0.0249	−0.0217
		(1.15)	(0.78)	(−0.35)
Easeofuse		0.808***	0.309***	0.648***
		(18.43)	(5.36)	(5.27)
Size		−0.675***	−0.265***	−0.529***
		(−23.49)	(−6.81)	(−5.49)
Videoquality		−1.626***	−0.695***	−0.604**
		(−20.66)	(−6.84)	(−3.17)
LCD		−1.783***	−0.694***	−1.074***
		(−40.27)	(−9.76)	(−6.40)
Design		−0.972***	−0.379**	−0.332
		(−10.09)	(−3.08)	(−1.18)
Batterylife		−0.809***	−0.301***	−0.253*
		(−27.50)	(−7.23)	(−2.55)
Log of *sales rank*			0.611***	0.462***
			(25.47)	(7.15)
_Cons	5.242***	4.476***	1.615***	2.905***
	(62.26)	(52.16)	(10.60)	(7.10)
N	7,307	7,307	7,267	1,349

Note. The *t*-statistics are in parentheses.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

• "The battery life is good (but *I wouldn't know what to compare it to*), but I would say it lasted me about a 1.5–2 days of shooting pics throughout the day (*I'm not sure what it should last*, but 1.5–2 days batt life is pretty good to me)."

• "The battery life is good, but *get a backup battery if you want to take a lot of pictures*."

• "The battery life is good *enough* to take on extended trips from your hotel."

• "Battery life is good. My personal camera is the Sony V1, it too is a nice camera, however the screen size is much smaller and the glossy finish makes is almost impossible to view in bright light. Overall both cameras are very nice. The DSCW-5 is quite a bit smaller and lighter than the V1. Another good feature about the DSCW5 is that is runs on two AA batteries that in an emergency you could purchase from any store. *Battery life however on non-rechargeable batteries would not be good, but it is better than having a dead specialized battery, like most other cameras have including my* V1."

Table 13 GMM and Dynamic GMM Estimation Results for Camcorder Category with Induced Polarity

	(1) Model 1 (GMM)	(2) Model 2 (GMM)	(3) Model 3 (DGMM)	(4) Model 4 (DGMM)
Price (unit = \$100)	0.0560***	0.0928***	0.0602***	0.0467*
	(3.55)	(6.23)	(4.03)	(2.11)
Trend	0.687***	0.734***	0.389***	0.432***
	(17.28)	(18.73)	(8.25)	(4.07)
Age	0.00688***	0.00656***	0.00371***	0.00318***
	(85.05)	(77.57)	(19.35)	(7.72)
Isused	0.341***	0.365***	0.201***	0.179***
	(17.42)	(18.87)	(8.75)	(3.75)
Fraction of one-star reviews	−3.362***	−3.323***	−1.849***	−1.525***
	(−28.84)	(−40.82)	(−14.53)	(−5.68)
Fraction of five-star reviews	1.746***	1.663***	0.927***	0.776***
	(21.61)	(28.25)	(11.84)	(5.08)
Numreviews (unit = 10)	−0.461***	−0.404***	−0.225***	−0.181***
	(−32.51)	(−28.14)	(−11.38)	(−4.01)
Rating	−1.017***	−1.105***	−0.643***	−0.577***
	(−47.93)	(−39.45)	(−15.06)	(−6.89)
Reviewlength (unit = 10,000 words)	0.122***	0.101***	0.0549***	0.0470
	(8.15)	(7.35)	(3.42)	(1.33)
Ratingstdev	−1.024***	−0.871***	−0.483***	−0.372***
	(−46.77)	(−44.41)	(−14.79)	(−5.51)
Hasrating	1.536***	2.209***	1.336***	1.346***
	(20.80)	(20.69)	(10.47)	(5.48)
Picturequality		−2.487***	−1.373***	−1.055***
		(−41.63)	(−14.09)	(−4.86)
Easeofuse		−0.526***	−0.268***	−0.235**
		(−14.32)	(−6.42)	(−2.99)
Weightsize		−0.547***	−0.252***	−0.260*
		(−8.03)	(−3.43)	(−2.02)
Videoquality		0.732***	0.391***	0.391***
		(19.73)	(8.45)	(4.23)
Log of *sales rank*			0.445***	0.561***
			(17.16)	(9.57)
_Cons	7.309***	6.689***	3.658***	2.765***
	(60.49)	(64.80)	(17.37)	(6.13)
N	4,377	4,377	4,356	790

Note. The *t*-statistics are in parentheses.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

That being noted, there is no doubt that we can demonstrate and claim that *review text has significant explanatory power for product sales* (*both as a contemporaneous indicator and as a forecasting tool*, as described in the next section). Causal interpretation of the results of our model should only be made while recognizing the potential noise and biases in the data.

4.4. Comparison of Methods

In this section, we discuss the strengths and weaknesses of the crowdsourcing-based solution versus the automated text-mining approach. We also discuss the strengths and weaknesses of the imposed polarity versus the inferred polarity approach. Such discussions would inform future researchers of the merits of each approach.

4.4.1. Crowdsourcing vs. Automatic Text Mining.

We can consider these approaches as solutions with different start-up costs and different variable costs.

Crowdsourcing has a low start-up cost, and any researcher can quickly use it for processing relatively large collections of text documents (e.g., a few thousand documents). The accuracy is good to great, and it can be done reasonably quickly. It is ideal for researchers that have a one-off task regarding content analysis and do not expect to repeatedly apply the same techniques for a variety of data sets. On the other hand, text mining has a much higher start-up cost. Setting up the algorithms requires significant expertise, and if the text analysis needs to be done in a new, previously unstudied domain, the development of the necessary resources requires both human

effort and time. However, it has the advantage of zero marginal cost once everything is set up. It can scale to collections with hundreds of thousands of documents and in many different data sets. Our analysis indicates that the analysis and noise level of the two approaches is similar. Therefore, the choice of which technique to use depends on "external" factors that are particular to each researcher.

4.4.2. Inferred vs. Imposed Polarity. The estimation with inferred polarity and the estimation with induced polarity represent two competing solutions to the same problem.

The imposed polarity approach takes a more "traditional" view of language. This approach assumes that language is static and has a given and unambiguous meaning. The main advantage of the imposed polarity is the reduction in the number of variables that are included in the econometric model. The imposed polarity approach brings outside actors to decide on the strength of particular evaluations and uses data only to evaluate feature weights. This efficiently solves the problem of data sparsity but potentially introduces human bias in the results.

The inferred polarity approach takes an agnostic view toward language. This model assumes that the way that humans interpret language (in this case, evaluations) depends on many contextual factors, and it is not possible to ex post assign a polarity and strength to the evaluations. Thus, the inferred polarity approach attempts to learn as much as possible about the language from the given data itself. However, the inferred polarity approach can only separate weights of individual features from strengths of individual evaluations under some assumptions. In addition, because the dimensionality of the data increases tremendously, we need to impose restrictions on what data can be included in the model.

We believe that the choice of which method to adopt depends on both the particular application chosen as well as the data set. For example, for a forecasting task, the inferred polarity model is likely to be preferred simply because it feeds more features into the machine-learning model (described above). We believe that the best way to interpret the coefficients of the inferred polarity model is from a purely predictive viewpoint. To separate weights of individual features from strengths of individual evaluations naturally, we recommend using the imposed polarity model.

5. Predictive Modeling of Sales Rank

In the previous section, we demonstrated that consumer opinions have a significant effect on product sales, and we can attempt to learn consumer preference for particular product features by relating changes in the product demand to the changes in the content of consumer reviews. In this section, we adopt a purely forecasting perspective and show that text of newly published product reviews can be used to predict short-term *future* changes to the product sales. In particular, we state our prediction task as follows: given the set of product reviews posted for a product within the last week and related changes to other numeric variables such as product price, predict

- whether the product sales (as measured by the sales rank) will go up or down within the next week,
- what the exact product sales rank after the next week will be.

5.1. Classification Problem

The first task is a binary classification task: We predict the sign of the value $SalesRank(t+7) - SalesRank(t)$ using whatever information is available at time t (measured in days). We have experimented with four different classifier types: logistic regression, support vector machines, decision trees, and random forests. Support vector machine slightly outperformed (2%–3%) logistic regression but took significantly longer time to train, whereas tree-based classifiers performed significantly worse on both product categories; thus, in the following, we report results of the logistic regression model.

For each of the three product categories, we estimated two models: the baseline model using all variables (including numeric review data) except for the review text and the full model, which additionally includes the top 20 most popular opinion phrases as features. For every feature, we included its absolute level at time t as well as its change within the past week ($Feature(t) - Feature(t-7)$).

Table 8 reports results of 10-fold cross-validation (Kohavi 1995) on each category based on the area under curve (AUC) metric. For digital cameras, we see that whereas the no-text model increases the AUC from the random baseline level of 0.5 to 0.574, the addition of textual content in the model raises the AUC from 0.574 to 0.644. In other words, we see an almost *twofold* increase in the AUC from the baseline case once we add the textual content of reviews. In the case of camcorders, we see an almost *threefold* increase in predictive power with text from the baseline case of 0.5 to 0.617, when compared to the AUC of the no-text model (0.544). Both these categories demonstrate substantial improvements in the AUC compared to the baseline cases when textual data are added to the models.

5.2. Regression Problem

The second task we consider is a regression problem, in which we try to predict the exact product sales rank one week into the future. This is somewhat similar

to the regression problem we studied in §3.2, however the independent variables are now lagged by one week, and we are only concerned about the out-of-sample predictive power of the model and not the structural interpretation of the coefficients.

In addition to the baseline model, which includes the top 20 opinion phrases for every product category, we also propose and evaluate a novel technique that allows including more regressors by making additional assumptions about the parameter space that reduce the dimensionality of the problem and avoid overfitting.

In §3.2, we defined $Score(f, e)$ of every opinion phrase (evaluation e applied to feature f) to be a joint function of the evaluation and the feature: The same evaluation applied to different features can have different relative impact, and vice versa, the same feature can have different relative weight when combined with different evaluations. For the predictive modeling part, we relax this assumption and assume that the score of each evaluation is independent of the feature being described. For example, the strength of the evaluation "great" (compared to the strength of other evaluations like "good") should be the same for both "picture quality" and "video quality" features. Formally, this can be written as

$$\exists S_F\colon \mathcal{F} \Rightarrow \mathbb{R},\ S_E\colon \mathcal{E} \Rightarrow \mathbb{R} \quad \forall f \in \mathcal{F},\ e \in \mathcal{E}$$
$$Score(f, e) = S_F(f) S_E(e).$$

In other words, every evaluation has certain weight $S_E(e)$ independent of the feature that it evaluates, and every feature has certain weight $S_F(f)$ independent of the evaluation applied to the feature; the impact of any particular opinion is calculated as a product of these two weights. For N features and M evaluations, the number of model parameters is therefore reduced from MN to $2M + N$.

Formally, the estimated model can be represented by the following equation:

$$\log(s_{jt+7}) - \log(s_{jt}) = \mathbf{x} \cdot \boldsymbol{\gamma} + \sum_{f \in \mathcal{F}} \sum_{e \in \mathcal{E}} Y_{jt}(f, e) \cdot S_E(e) \cdot S_F(f) + \varepsilon_{jt}, \qquad (5)$$

where s_{jt} represents product j sales rank at time t, vector $(Y_{jt}t(f, e))_{f \in \mathcal{F}, e \in \mathcal{E}}$ represents review opinions available at time t, and vector $\mathbf{x}$ represents all other numeric variables such as product price.

Although the model is nonlinear, it can be estimated by a sequence of regression operations. We use the observation that, for a fixed vector S_E, Equation (5) represents a regression for S_F. Vice versa, for a fixed vector S_F, Equation (5) represents a regression for S_E. Because the model exhibits significant nonlinearity in parameters and can potentially overfit the data, we further add a regularization term $\lambda \cdot \|S_E\|^2 \cdot \|S_F\|^2$ to the optimization function. Overall, the estimation algorithm is as follows:

$S_F(f) \equiv 1$ {all features are initially assumed to be equally important}
while not converged **do**
  $S_E(e) \Leftarrow$ coefficients from regression of $\log(r_{jt+7}) - \log(r_{jt})$ on $\mathbf{x}$ and $S_F(f) w^t(f, e)$ with regularization weight $\lambda \|S_F\|^2$
  $S_F(f) \Leftarrow$ coefficients from regression of $\log(r_{jt+7}) - \log(r_{jt})$ on $\mathbf{x}$ and $S_E(e) w^t(f, e)$ with regularization weight $\lambda \|S_E\|^2$
end while

We have implemented the algorithm above with 10 different feature names and 10 different evaluation adjectives (i.e., 100 different opinion phrases) in both product categories and compared its results with results of a simple regression model including only the top 20 most popular opinion phrases. In every category, we saw a 5% to 10% increase in predictive power as measured by out-of-sample R^2.

6. Managerial Implications and Conclusions

We are the first to combine a theoretically motivated econometric model with text-mining techniques to study the influence of textual product reviews on product choice decisions. Using a unique data set from a leading online retailer of electronic products, Amazon.com, we demonstrate the value of combining textual data with econometric and predictive modeling for quantitative interpretation of user-generated content. Our empirical approach is able to impute which product features described in reviews are more important to consumers and how one can quantify opinions contained in the textual component of reviews. The results of our study indicate that the textual content in product reviews has a significant predictive power for consumer behavior and explains a large part of the variation in product demand over and above the impact of changes in numeric information such as product price, product age, trends, seasonal effects, and the valence and the volume of reviews.

Our results have several managerial implications. Most consumer products have a mix of attributes that can be objectively evaluated prior to purchase and subjective attributes that are harder to quantitatively evaluate. In this vein, this distinction between subjective and objective attributes is similar to the distinction between search and experience goods (Nelson 1970). One of the interesting applications of our text-based approach is that it allows us to easily incorporate, into quantitative models, product attributes

that were inherently qualitative and hence difficult to measure and incorporate into econometric models. Using our approach, it is now possible to infer the weight that customers put in such features. For example, in the case of digital cameras, attributes like "design" and "ease of use" are attributes that are hard to evaluate quantitatively because they are susceptible to subjectivity of the evaluator. On the other hand, attributes like "battery life" and size would belong to the more "objective features" category. Attributes like "picture quality" would fall somewhere in the middle. Our focus on combining econometrics with automated text-based analyses can help recover the relative economic weight that consumers place on these features, irrespective of whether they are "objective" of "subjective." Our technique can be of interest to manufacturers and retailers to determine which features contribute the most to the sales of their products. Such information, for example, can help manufacturers facilitate changes in product design over the course of a product's life cycle as well as help retailers decide on which features to promote and highlight in advertisements and in-store displays.

Our paper also provides some insights to online advertisers (manufacturers or retailers) who aim to use customer-generated opinions to automatically devise an online advertising strategy for each product using the widely popular model of sponsored search advertising. For instance, our methods can be extrapolated to different product categories for firms to select the appropriate keywords to bid in these advertising auctions, and for selecting the most pertinent text in the advertisement that highlights the differentiating characteristics of the advertised products that consumers value the most. For example, if the phrase "excellent video quality" is associated with increase in sales three times more than the phrase "great design" for a given model of Sony digital cameras, then it might make sense for the manufacturer or the retailer to choose the set of keywords associated with the former phrase rather than the latter.

We would like to note that methodologies presented in this paper possess flexibility: Although some of our current choices can be derived from a simple model of Bayesian learning by consumers, there are alternative approaches for almost all steps of the text-mining process, from feature extraction to choosing a particular functional form for the estimation equation. To the best of our knowledge, our paper is the first application of text mining to demand estimation and it provides encouraging results by showing that even a simple choice model combined with simple text-mining techniques can have significant explanatory power. Overall, we believe that the interaction of economics and marketing mix models with text-mining tools from natural-language processing can benefit both fields. Economic approaches can offer natural solutions to text-mining problems that seemed too hard to solve in a vacuum (e.g., determining the strength of an opinion). Similarly, text-mining approaches can improve the current state of the art in empirical economics, where the focus has traditionally been on relatively small, numeric data sets.

Although we have taken a first step in several directions, we acknowledge that our approach has several limitations, some borne by the nature of the data themselves. Our work attempts to combine econometric modeling with text-mining techniques and can benefit from parallel improvements in both fields. In particular, the methodologies presented in this paper can benefit from improvements in discrete choice modeling and in text mining. Better techniques for handling absence of individual-level data, overcoming sparsity of textual review contents, improvements on natural-language processing algorithms, and better techniques for handling noisy information on product sales would all result from improvements of our own work.

Our research has certain limitations. First, our methods and results are better suited for vertically differentiated products like electronics. Future work could examine what kind of empirical models could be applied products that are horizontally differentiated. Second, our approach implicitly assumes that consumers learn independently across attributes and independently for each product. In reality, consumers may learn about quality levels across attributes or even across products. Consumer might even engage into a process of learning the *weights* that they should place in each product feature by reading the reviews of other, more experiences customers. Application of more advanced models of learning with uncertainty can potentially provide better insights. Third, some of the variables in our data are proxies for the actual variables needed for more advanced empirical modeling. For example, we use sales rank as a proxy for demand (Brynjolfsson et al. 2003, Chevalier and Goolsbee 2003, Ghose and Sundararajan 2006) from one retailer. Future work can use real demand data from multiple retailers for estimating the value of different product features, as in the paper by Ghose et al. (2011), who estimate demand for hotels using actual transactions. To control for effects of potential advertising shocks and the inherent endogeneity in the word of mouth–sales relationship, we used Google Trends data as a measure of publicity. Future work can incorporate more explicit and better measures of publicity and advertising such as in Luan and Neslin (2009). Notwithstanding these limitations, we hope our paper paves the way for future research in this exciting domain.

Acknowledgments

The authors thank Rhong Zheng for assistance with data collection. They offer deep thanks to Kenneth Reisman and his company Pluribo.com for providing them with an ontology construction tool. They thank seminar participants at Microsoft Research, IBM Research, Yahoo Research, Carnegie Mellon University, Columbia University, New York University, University of Utah, SCECR 2008, INFORMS-CIST 2008, and the 2008 NET Institute Conference for helpful comments. This work was supported by a 2006 Microsoft Live Labs Search Award, a 2007 Microsoft Virtual Earth Award, and by National Science Foundation CAREER Grants IIS-0643847 and IIS-0643846. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the Microsoft Corporation or of the National Science Foundation.

Appendix. Figures, Summary Statistics, and Estimation Results

Theoretical Model

In this appendix, we present a justification of our empirical approach based on a simple theoretical model of multiattribute choice under uncertainty. Although the model is not required to understand the methodologies and the results of this paper, through the description of the model, we hope to outline clearly the scope and applicability of our research, explain what are the implicit assumptions behind our current approach are, and identify directions for future research. The foundation of our model is the seminal paper of Roberts and Urban (1988). Products have multiple attributes, and the quality of each attribute is uncertain to consumers. To reduce uncertainty, consumers read product reviews and use Bayesian learning to update their beliefs about the quality of product attributes. Based on their beliefs, consumers buy the product that maximizes their expected utility, a fact that is reflected in the product sales. We outline the basic concepts of the model below.

Multiattribute Products. We model products as n-dimensional vectors of well-defined product attributes. Ignoring the uncertainty aspect, our model will represent every product by an n-dimensional point $z_j = (z_{1j}, \ldots, z_{nj})$, where each z_{ij} should be read as the amount or quality of the ith attribute for the jth good. Although natural in many markets, such as markets for consumer appliances, this assumption indicates that our model cannot be applied to products such as movies or music that cannot be represented by a small set of well-defined attributes.

Preferences. We assume a simple scenario of homogeneous preferences for product attributes. To incorporate risk aversion in our model, we abstain from linear utility setting, instead adopting negative exponential utility (Roberts and Urban 1988, Bell and Raiffa 1988).[9] Formally, we assume that for any consumer and any product $\bar{z}_j$ with *deterministic* vector of attributes $(z_{1j}, \ldots, z_{nj})$ and price p_j, the utility of purchasing the product is given by

$$u(\bar{z}_j) = -\exp\left(\alpha p_j - \sum_{k=1}^{n} \beta_k z_{kj} + \varepsilon_{ij}\right), \tag{6}$$

where ε_{ij} is the "taste for randomness" residual representing inherent randomness in consumer's choice process.

Uncertainty. Instead of having a direct assessment of vector $\bar{z}_j$ for each product, consumers are uncertain and have beliefs about the distribution of $\bar{z}_j$. We futher assume that consumers share a common information set and therefore have common beliefs about the attributes of the jth product represented by the distribution function F_j. In such a scenario, consumers making purchase decisions are not choosing just a product (i.e., a bundle of attributes), but they choose a lottery over bundles of attributes. We follow classic modeling approach for choice under uncertainty and adopt the von Neumann–Morgenstern expected utility framework: consumers always choose product $\bar{z}_j$ with the highest expected utility $Eu(\bar{z}_j)$.

Prior Beliefs. Application of our theory requires specification of the form of consumers' beliefs. We assume normal prior beliefs with a diagonal covariance matrix:[10]

$$F_j \sim \mathcal{N}\left(\begin{bmatrix}\mu_{1j}\\ \mu_{2j}\\ \cdots\\ \mu_{nj}\end{bmatrix}, \begin{bmatrix}\sigma_{1j}^2 & 0 & 0 & \cdots\\ 0 & \sigma_{2j}^2 & 0 & \cdots\\ \cdots & \cdots & \cdots & \cdots\\ \cdots & \cdots & 0 & \sigma_{nj}^2\end{bmatrix}\right).$$

Our argument is that, under certain regularity conditions, recursive the Bayesian learning process results in asymptotic normality of the posterior distribution (Chen 1985). Consumers often use their previous experiences with similar products to form a prior distribution about the quality of a new product. Because such experiences correspond to a type of recursive Bayesian learning process, we can assume that consumers use normal priors. This is also consistent with the original approach of Roberts and Urban (1988), who also assumed that consumers' uncertainty is characterized by a normal distribution. It can be shown (Roberts and Urban 1988) that, in combination with negative exponential utility, it gives a particularly simple analytic representation of the expected utility function:

$$Eu(\bar{z}_j) = -\exp\left(\alpha p_j - \sum_{k=1}^{n} \beta_k \mu_{jk} + \tfrac{1}{2}\sum_{k=1}^{n} \beta_k^2 \sigma_{jk}^2 + \varepsilon_{ij}\right). \tag{7}$$

It immediately follows that consumers will prefer product $\bar{z}_j$ to product $\bar{z}_l$ if and only if

$$\varepsilon_{ij} - \varepsilon_{il} \geq -\alpha(p_j - p_l) + \sum_{k=1}^{n} \beta_k(\mu_{jk} - \mu_{lk}) - \tfrac{1}{2}\sum_{k=1}^{n} \beta_k^2(\sigma_{jk}^2 - \sigma_{lk}^2), \tag{8}$$

[9] Roberts and Urban (1988) provide an elaborate argument in favor of negative exponential utility based on the observation that, for measurable value functions, if the consumer obeys the von Neumann–Morgenstern axioms for lotteries, and if a utility function exists, the value function should show constant risk aversion with respect to the strength of preference measure (Bell and Raiffa 1988).

[10] Diagonality of the covariance matrix implicitly enforces independence of beliefs for different attributes. For example, additional information about picture quality for a digital camera should not affect consumers' beliefs about its battery life.

or

$$\varepsilon_{ij} - \varepsilon_{il} \geq (-\alpha p_j + R(F_j)) - (-\alpha p_l + R(F_l)), \quad (9)$$

where

$$R(F_j) = \sum_{k=1}^{n} \beta_k \mu_{jk} - \tfrac{1}{2} \sum_{k=1}^{n} \beta_k^2 \sigma_{jk}^2 \quad (10)$$

is the so-called "risk-adjusted preference function" (Roberts and Urban 1988) representing the expected value of the product attributes after discounting for the uncertainty associated with the product.

Idiosyncratic Preferences. We further follow Roberts and Urban (1988) and assume that the "taste for randomness" term ε_{ij} is uncorrelated to particular product attributes and follows the type I extreme value distribution. This assumption gives a familiar logit expression for the probability P_j that consumers choose the product j,

$$P_j = \frac{\exp(-\alpha p_j + R(F_j))}{\sum_l \exp(-\alpha p_l + R(F_l))}, \quad (11)$$

and, assuming that the total mass of consumers is normalized to one, a similar expression for the expected product demand,

$$s_j = \frac{\exp(-\alpha p_j + R(F_j))}{\sum_l \exp(-\alpha p_l + R(F_l))}. \quad (12)$$

We are now ready to formulate our main result.

Lemma 1 (First-Order Effects of Changes in Beliefs). *Assume that in Equation* (12)*, the "risk-adjusted preference function" for product j is changed by $\triangle R$, i.e., $\hat{R}(F_j) = R(F_j) + \triangle R$ and $\hat{R}(F_l) = R(F_l)$ for $l \neq j$. Let $\hat{s}_j$ represents the new market share for the product j. Then,*

$$\log(\bar{s}_j) - \log(s_j) = \triangle R + \epsilon, \quad (13)$$

where

$$|\epsilon| \leq \frac{|\bar{s}_j \triangle R|}{1 - |\bar{s}_j \triangle R|} = o(|\triangle R|). \quad (14)$$

Proof. We will proceed assuming $\triangle R \geq 0$; the other case follows by symmetry. Define $Z = \sum_{l \neq j} \exp(-\alpha p_l + R(F_l))$;

$$\bar{s}_j = \frac{\exp(-\alpha p_j + R(F_j) + \triangle R)}{\exp(-\alpha p_j + R(F_j) + \triangle R) + Z}$$

$$= \exp(\triangle R) \frac{\exp(-\alpha p_j + R(F_j))}{\exp(-\alpha p_j + R(F_j) + \triangle R) + Z} \quad (15)$$

$$= \exp(\triangle R) \frac{\exp(-\alpha p_j + R(F_j))}{\exp(-\alpha p_j + R(F_j)) + Z} \cdot \frac{\exp(-\alpha p_j + R(F_j)) + Z}{\exp(-\alpha p_j + R(F_j) + \triangle R) + Z} \quad (16)$$

$$= \exp(\triangle R) s_j U, \quad (17)$$

where

$$U = \frac{\exp(-\alpha p_j + R(F_j)) + Z}{\exp(-\alpha p_j + R(F_j) + \triangle R) + Z}. \quad (18)$$

After taking logs,

$$\log(\bar{s}_j) - \log(s_j) = \triangle R + \log(U), \quad (19)$$

thus Equation (13) holds with $\epsilon \equiv \log(U)$. It remains to put the bound on ϵ. Using well-known inequality $(1 - e^{-x}) \leq x$ for $x \geq 0$, one can show that

$$|1 - U| = \left| \frac{\exp(-\alpha p_j + R(F_j) + \triangle R) - \exp(-\alpha p_j + R(F_j))}{\exp(-\alpha p_j + R(F_j) + \triangle R) + Z} \right| \quad (20)$$

$$= |\bar{s}_j (1 - \exp(-\triangle R))| \leq |\bar{s}_j \triangle R|. \quad (21)$$

Also, $\log(U) \leq (1 - U)/U$ for $0 < U \leq 1$; therefore,

$$\log(U) \leq \frac{1 - U}{U} = \frac{1 - U}{1 - (1 - U)} \leq \frac{|\bar{s}_j \triangle R|}{1 - |\bar{s}_j \triangle R|}. \quad \text{Q.E.D.} \quad (22)$$

The main message of Lemma 1 is that the effect of changes in the risk-adjusted preference function on the product sales can be approximated by a linear function, and unless the product under consideration controls a significant fraction of the total market share, the error of such approximation is negligible for practical purposes.

Bayesian Updating of Beliefs. The final component of our model is the mechanism used by consumers to update their beliefs. We use a Bayesian learning approach. For simplicity of estimation, we assume that the qualities of different product attributes are learned independently. For example, observing the picture quality of a digital camera does not give consumers much information on the camera design, the camera size, battery life, etc. Note that this assumption can hold either because signals of different features are actually independent or because consumers have bounded rationality and cannot capture complex dependencies. Although a number of marketing models allow for cross-attribute consumer learning, e.g., Bradlow et al. (2004), because of the limited size of our data set, we leave such extensions as directions for future research.

A convenient choice of the likelihood function is the conjugate distribution, which, as we assume normal priors, is also normal. In this setting, if consumers current beliefs about quality of the kth attribute for the product j are given by the distribution $N(\mu_{kj}, \sigma_{kj}^2)$, the variance of the likelihood function is τ_k^2, and consumers observe a sequence of signals $\{x_{kj1}, \ldots, x_{kjm_{jk}}\}$, then the posterior distribution of beliefs about the kth product attribute will be

$$N\left(\mu_{kj} + \frac{1}{m_{kj} + \eta_{kj}} \sum_{r=1}^{m_{kj}} (x_{kjr} - \mu_{kj}), \frac{1}{m_{kj} + \eta_{kj}} \sigma_{kj}^2\right), \quad (23)$$

where $\eta_{kj} = \tau_k^2 / \sigma_{kj}^2$ represents the strength of prior beliefs about the attribute quality (Duda et al. 2000).

In particular, with the negative exponential utility assumption, the risk-adjusted preference function is

$$R(F_j) = \sum_{k=1}^{K} \beta_k \frac{1}{m_{kj} + \eta_{kj}} \sum_{r=1}^{m_{kj}} (x_{kjr} - \mu_{kj}) - \frac{1}{2} \beta_k^2 \frac{1}{m_{kj} + \eta_{kj}} \sigma_{kj}^2. \quad (24)$$

Connection to the Estimation Equation. Our resulting estimation Equation (2) is obtained by connecting Equations (13) and (24) and adopting the following conventions:

- Every feature $f \in \mathcal{F}$ described in product reviews represents one product dimension k.
- Every evaluation $e \in \mathcal{E}$ for a particular feature f represents a single quality signal x_{kjr}. There is the following relationship between signal x_{kjr} and opinion weight $Score(f, e)$:

$$Score(f, e) = \beta_k (x_{kjr} - \mu_{kj}). \quad (25)$$

- The weighting coefficients for opinions are defined as

$$Y_{jt}(f,e) = \frac{N(f,e)}{m_{kj} + \eta_{kj}} \tag{26}$$

if opinion phrase (f,e) is mentioned $N(f,e)$ times in the data, and m_{kj} is the total number of opinions for feature f in product reviews for product j (both are measured as of time t); η_{kj} is represented by the smoothing factor s in Equation (3).

- We do not directly incorporate the variance terms σ_{kj}^2 in the estimation equation because of the absence of a non-ambiguous approach to measure them. Instead, we control for variance of consumer beliefs by including a number of control variables in the model, such as the number of reviews, the fraction of one- and five-star reviews, and the standard deviation of numeric review ratings.

We would like to conclude by noting that Equation (25) provides interesting identification insights. It shows that, without making additional assumptions and obtaining additional data, we cannot separate the following three effects: the effect of the information contained in the opinion or opinion strength (x_{kjr}), the prior consumers' expectations of the feature quality (μ_{kj}), and the sensitivity of the consumers' utility to the particular feature (β_k).

References

Arellano, M., S. Bond. 1991. Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *Rev. Econom. Stud.* **58**(2) 277–297.

Arellano, M., O. Bover. 1995. Another look at the instrumental variable estimation of error-components models. *J. Econometrics* **68**(1) 29–51.

Bell, D., H. Raiffa. 1988. Managerial value and intrinsic risk aversion. *Decision Making: Descriptive, Normative, and Prescriptive Interactions*. Cambridge University Press, New York, 384–397.

Berry, S., J. Levinsohn, A. Pakes. 1995. Automobile prices in market equilibrium. *Econometrica* **63**(4) 841–890.

Bickart, B., R. M. Schindler. 2001. Internet forums as influential sources of consumer information. *J. Interactive Marketing* **15**(3) 31–40.

Blundell, R., S. Bond. 1998. Initial conditions and moment restrictions in dynamic panel data models. *J. Econometrics* **87**(1) 115–143.

Bradlow, E. T., Y. Hu, T.-H. Ho. 2004. A learning-based model for imputing missing levels in partial conjoint profiles. *J. Marketing Res.* **41**(4) 369–381.

Brynjolfsson, E., Y. Hu, M. Smith. 2003. Consumer surplus in the digital economy: Estimating the value of increased product variety. *Management Sci.* **49**(11) 1580–1596.

Chen, C.-F. 1985. On asymptotic normality of limiting density functions with bayesian implications. *J. Roy. Statist. Soc.* **47**(3) 540–546.

Chen, Y., J. Xie. 2008. Online consumer review: Word-of-mouth as a new element of marketing communication mix. *Management Sci.* **54**(3) 477–491.

Chevalier, J. A., A. Goolsbee. 2003. Measuring prices and price competition online: Amazon.com and BarnesandNoble.com. *Quant. Marketing Econom.* **1**(2) 203–222.

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

Das, S. R., M. Chen. 2007. Yahoo! for Amazon: Sentiment extraction from small talk on the Web. *Management Sci.* **53**(9) 1375–1388.

Decker, R., M. Trusov. 2010. Estimating aggregate consumer preferences from online product reviews. *Internat. J. Res. Marketing* **27**(4) 293–307.

Dellarocas, C., N. Farag Awady, X. (Michael) Zhang. 2007. Exploring the value of online product ratings in revenue forecasting: The case of motion pictures. Working paper, Robert H. Smith School of Business, University of Maryland, College Park.

Duan, W., B. Gu, A. B. Whinston. 2005. Do online reviews matter? An empirical investigation of panel data. Technical report, McCombs Research Paper Series, University of Texas at Austin, Austin.

Duda, R. O., P. E. Hart, D. G. Stork. 2000. *Pattern Classification*, 2nd ed. John Wiley & Sons, New York.

Eliashberg, J., S. K. Hui, Z. J. Zhang. 2007. From story line to box office: A new approach for green-lighting movie scripts. *Management Sci.* **53**(6) 881–893.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Forman, C., A. Ghose, B. Wiesenfeld. 2008. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Inform. Systems Res.* **19**(3) 291–313.

Frenkel, T. H., Y. Kim, M. Wedel. 2002. Bayesian prediction in hybrid conjoint analysis. *J. Marketing Res.* **39**(2) 253–261.

Ghani, R., K. Probst, Y. Liu, M. Krema, A. Fano. 2006. Text mining for product attribute extraction. *SIGKDD Explorations* **1**(8) 41–48.

Ghose, A., P. G. Ipeirotis. 2010. Estimating the helpfulness and economic inpact of product reviews: Mining text and reviewer characteristics. *IEEE Trans. Knowledge Data Engrg.*, IEEE Computer Society, Washington, DC. http//doi.ieeecomputersociety.org/10.1109/TKDE.2010.188.

Ghose, A., P. G. Ipeirotis, A. Sundararajan. 2007. Opinion mining using econometrics: A case study on reputation systems. *Proc. 44th Annual Meeting Assoc. Comput. Linguistics* (*ACL 2007*), Association for Computational Linguistics, Stroudsburg, PA, 416–423.

Ghose, A., P. G. Ipeirotis, B. Li. 2011. Designing ranking systems for hotels on travel search engines by mining user-generated and crowd-sourced content. Working paper, New York University, New York.

Ghose, A., A. Sundararajan. 2006. Evaluating pricing strategy using ecommerce data: Evidence and estimation challenges. *Statist. Sci.* **21**(2) 131–142.

Gilbride, T. J., P. J. Lenk, J. D. Brazell. 2008. Market share constraints and the loss function in choice-based conjoint analysis. *Marketing Sci.* **27**(6) 995–1011.

Godes, D., D. Mayzlin. 2004. Using online conversations to study word-of-mouth communication. *Marketing Sci.* **23**(4) 545–560.

Green, P., V. Srinivasan. 1978. Conjoint analysis in consumer research: Issues and outlook. *J. Consumer Res.* **5**(2) 103–123.

Hansen, L. P. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* **50**(4) 1029–1054.

Horsky, D., S. Misra, P. Nelson. 2006. Observed and unobserved preference heterogeneity in brand-choice models. *Marketing Sci.* **25**(4) 322–335.

Hu, M., B. Liu. 2004. Mining and summarizing customer reviews. *Proc. 10th ACM SIGKDD Internat. Conf. Knowledge Discovery and Data Mining* (*KDD-2004*), Association for Computing Machinery, New York, 168–177.

Hu, N., P. Pavlou, J. Zhang. 2008. Why do online product reviews have a J-shaped distribution? Overcoming biases in online word-of-mouth communication. Working paper, Singapore Management University, Singapore.

Johnson, R. 1987. Adaptive conjoint analysis. *Sawtooth Software Conf. Perceptual Mapping, Conjoint Anal. Comput. Interviewing*, Sawtooth Software, Sequim, WA.

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proc. 14th Internat. Joint Conf. Artificial Intelligence* (*IJCAI-95*), American Association for Artificial Intelligence, Menlo Park, CA, 1137–1143.

Lee, T., E. Bradlow. 2007. Automatic construction of conjoint attributes and levels from online customer reviews. Working paper, The Wharton School, University of Pennsylvania, Philadelphia.

Li, X., L. M. Hitt. 2008. Self-selection and information role of online product reviews. *Inform. Systems Res.* **19**(4) 456–474.

Liu, B., M. Hu, J. Cheng. 2005. Opinion observer: Analyzing and comparing opinions on the Web. *Proc. 14th Internat. World Wide Web Conf.* (*WWW 2005*), *Chiba, Japan*, 342–351.

Liu, Y. 2006. Word of mouth for movies: Its dynamics and impact on box office revenue. *J. Marketing* **70**(3) 74–89.

Luan, J., S. Neslin. 2009. The development and impact of consumer word of mouth in new product diffusion. Working paper, Tuck School of Business at Dartmouth, Hanover, NH.

Manning, C. D., H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Marshall, P., E. Bradlow. 2002. A unified approach to conjoint analysis models. *J. Amer. Statist. Assoc.* **97**(459) 674–682.

Moulton, B. R. 1986. Random group effects and the precision of regression estimates. *J. Econometrics* **32**(3) 385–397.

Nelson, P. 1970. Information and consumer behavior. *J. Political Econom.* **78**(2) 311–329.

Netzer, O., R. Feldman, M. Fresko, J. Goldenberg. 2011. Mine your own business:market structure surveillance through text mining. Working paper, Columbia University, New York.

Pakes, A. 2003. A reconsideration of hedonic price indexes with an application to PCs. *Amer. Econom. Rev.* **93**(5) 1578–1596.

Pang, B., L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends Inform. Retrieval* **2**(1–2) 1–135.

Roberts, J. H., G. L. Urban. 1988. Modeling multiattribute utility, risk, and belief dynamics for new consumer durable brand choice. *Management Sci.* **34**(2) 167–185.

Roodman, D. 2006. How to do xtabond2: An introduction to "difference" and "system" GMM in Stata. Working Paper 103, Center for Global Development, Washington, DC.

Rosen, S. 1974. Hedonic prices and implicit markets: Product differentiation in pure competition. *J. Political Econom.* **82**(1) 34–55.

Sargan, J. D. 1958. The estimation of economic relationships using instrumental variables. *Econometrica* **26**(3) 393–415.

Sheng, V. S., F. Provost, P. Ipeirotis. 2008. Get another label? Improving data quality and data mining using multiple, noisy labelers. *Proc. 14th ACM SIGKDD Internat. Conf. Knowledge Discovery and Data Mining* (*KDD-2007*), Association for Computing Machinery, New York, 614–622.

Snow, R., B. O'Connor, D. Jurafsky, A. Y. Ng. 2008. Cheap and fast but is it good? evaluating non-expert annotations for natural language tasks. *Proc. Conf. Empirical Methods in Natural Language Processing* (*EMNLP 2008*), Association for Computational Linguistics, Stroudsburg, PA, 254–263.

Srinivasan, V. 1988. A conjunctive-compensatory approach to the self-explication of multiattributed preferences. *Decision Sci.* **19**(2) 295–305.

Toubia, O., D. I. Simester, J. R. Hauser, E. Dahan. 2003. Fast polyhedral adaptive conjoint estimation. *Marketing Sci.* **22**(3) 273–303.

Turney, P. D. 2002. Thumbs up or thumbs down? Semactic orientation applied to unsupervised classification of reviews. *Proc. 40th Annual Meeting Assoc. Comput. Linguistics* (*ACL 2002*), Association for Computational Linguistics, Stroudsburg, PA, 417–424.

Turney, P. D., M. L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inform. Systems* **21**(4) 315–346.

Villas-Boas, J. M., R. S. Winer. 1999. Endogeneity in brand choice models. *Management Sci.* **45**(10) 1324–1338.

von Ahn, L., L. Dabbish. 2004. Labeling images with a computer game. *CHI '04: Proc. SIGCHI Conf. Human Factors Comput. Systems*, Association for Computing Machinery, New York, 319–326.

Windmeijer, F. 2005. A finite sample correction for the variance of linear efficient two-step GMM estimators. *J. Econometrics* **126**(1) 25–51.