The Economic Impact of User-Generated Content on the Internet: Combining Text Mining with Demand Estimation in the Hotel Industry

Anindya Ghose
New York University
aghose@stern.nyu.edu

Panagiotis Ipeirotis
New York University
panos@stern.nyu.edu

Beibei Li
New York University
bli@stern.nyu.edu

September 2009

Abstract

Increasingly, user-generated product reviews, images and tags serve as a valuable source of information for customers making product choices online. An extant stream of work has looked at the economic impact of reviews. Typically, the impact of product reviews has been incorporated by numeric variables representing the valence and volume of reviews. In this paper, we posit that the information embedded in product reviews cannot be fully 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. Based on a unique dataset of hotel reservations available to us from Travelocity, we estimate demand for hotels using a two-step random coefficient based structural model. We use text mining techniques that allow us to incorporate textual information from user review in demand estimation models by inferring the sentiments embedded in them and supplement them with image classification techniques. The dataset contains complete information on transactions conducted over a 3 month period from Nov – Jan 2009 for hotels in the US. We have data on user- generated content from three sources: (i) user-generated hotel reviews from two well known travel search engines, Travelocity and Tripadvisor, (ii) tags generated by users identifying different locational attributes of hotels from Geonames.org, and (iii) user contributed opinions on the most important hotel characteristics from Amazon Mechanical Turk. Moreover, since some location-based characteristics, such as proximity to the beach, are not directly measurable based on UGC, we use image classification techniques to infer such features from the satellite images of the area. These different data sources are then merged to create one comprehensive dataset that enables us to estimate the weight that consumers place on different hotel characteristics. We then propose to design a new hotel ranking and recommendation system based on the empirical estimates of consumer surplus from hotel transactions. By improving the recommendation strategy of travel search engines, it can raise the conversion rate for a particular hotel, hence increasing the return-on-investment for travel search engines.

Keywords: Structural Modeling, User-Generated Content, Demand Estimation, Hotel Search, Social Media, Image Classification

TARGET JOURNAL: *MARKETING SCIENCE* (TO BE SUBMITTED BY DECEMBER 2009)[1]

[1] **BY INVITATION BASED ON A WHARTON-MSI GRANT COMPETITION ON USER-GENERATED CONTENT**

1. Introduction

The growing pervasiveness of the Internet has changed the way that consumers shop for goods. While 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. It comes as no surprise that online shoppers increasingly rely on alternative sources of information such as "word of mouth," in general and user-generated product reviews, in particular. In contrast to product descriptions provided by vendors, consumer reviews are, by construction, more user-oriented: in a review, customers describe a product in terms of usage scenarios and evaluate the product from a user's perspective (Chen and Xie 2004). 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 (for example, Chevalier and Mayzlin 2006, Clemons et al. 2006, Dellarocas et al. 2007, Duan et al. 2008, Forman et al. 2008, Godes and Mayzlin 2004). However, these studies have only used the numeric review ratings (e.g., the valence and the volume of reviews) in their empirical analysis, without formally incorporating the information contained in the text of the reviews. 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. (2006) estimate the impact of buyer textual feedback on price premiums charged by sellers in online second-hand markets. Eliashberg et al. (2007) combines natural-language processing techniques, and statistical learning methods to forecast a movies return on investment based only on textual information available in movie scripts. Archak et al. (2008) identify the weight that consumers put on individual evaluations and product features by estimating the impact of review text on sales using Amazon data. Ghose and Ipeirotis (2008) analyze the socio/economic impact of users' online product reviews and find that three broad feature categories influence user helpfulness – reviewer-related features, review subjectivity features, and review readability features. Pavlou and Dimoka (2006) and Ghose (2009) use content analysis techniques to mine the buyer-generated textual feedback of seller reputations, and examine the role of product uncertainty and seller uncertainty in influencing adverse selection in online used-good markets. But these studies typically do not focus on estimating the *impact of user-generated product reviews in influencing sales beyond the effect of numeric review ratings* which is one of the key research objectives of this paper.

There is another 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, while 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. 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. Further, 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. Therefore, our second research objective in this paper is to analyze the extent to which user generated reviews can help us learn consumer preferences for different product or service (in our case, hotel) attributes, both internal and external.

We undertake this study in the context of actual demand for hotel rooms using a unique dataset consisting of actual transactions and different kind of UGC, which we describe later. It is now widely acknowledged that local search for hotel accommodations is a component of general Web searches that is increasing in popularity as more and more users search for prices and reserve their trips online. Consumers, who are increasingly better informed, are becoming more demanding in the online tourism world. Customers try to identify hotels that satisfy particular criteria, such as food quality, service, locational attributes, and so on. Furthermore, given the recommended hotel and its price, customers would typically like to find out whether or not they are being overcharged in comparison to the "real value" of that hotel. Hence, locating a hotel with specific desired characteristics but without compromising on the value becomes an important question. Online travel search engines provide only rudimentary ranking facilities, typically using a single ranking criterion such as distance from the city center, star ratings, price per night, etc. This approach has quite a few shortcomings. First, it ignores the multidimensional preferences of the consumer in that a customer's ideal choice may consist of several hotel-specific attributes. Second, it largely ignores characteristics related to the location of the hotel, for instance, in terms of proximity to the beach or proximity to a downtown shopping area. These location-based features represent important characteristics that can influence the desirability of a particular hotel.

Currently there are no established metrics that can isolate the importance of the different characteristics of the hotels. The lack of authority and standardization in hotel ratings systems makes it that much harder for users to credibly infer the actual value from staying in a hotel. Existing empirical

work has only focused on 1-2 location-based characteristics for very small number of (usually 10-20) hotels within small geographical areas (White 2000, Bull 1998). This is where the emerging phenomenon of UGC can be really useful for researchers interested in inferring characteristics of hotels that matter most to consumers. Using demand estimation techniques, we propose to estimate the weight that consumers place on different hotel characteristics.

More specifically, we estimate demand for hotels as a function of its internal (service) and external (locational) characteristics, and thereby generate a general hotel ranking system that can serve the general customer population. We have access to a unique dataset of hotel reservations from Travelocity. The dataset contains complete information on transactions conducted over a 3 month period from Nov 2008 – Jan 2009 for 2117 hotels in the US. We have collected data on user- generated content from three sources: (i) user-generated hotel reviews from two well known travel search engines, Travelocity and Tripadvisor, (ii) tags generated by users identifying different locational attributes of hotels from Geonames.org, and (iii) user contributed opinions on the most important hotel characteristics from Amazon Mechanical Turk. Moreover, since some location-based characteristics, such as proximity to the beach, are not directly measurable based on UGC, we use image classification techniques to infer such features from the satellite images of the area. These different data sources are then merged to create one comprehensive dataset that enables us to quantify the economic value of UGC on the Internet. Using demand estimation techniques, we then aim to estimate the weight that consumers place on different hotel characteristics. The final outcome of our analysis allows us to compute the "value for the money" of a particular hotel based on estimation of price elasticities and consumer surplus. Thereafter we can generate hotel rankings that are superior to existing techniques seen in travel-related search engines.

To summarize, we combine structural modeling with text mining of user-generated reviews, analysis of social geotagging websites, and image classification methods. Our work involves three stages:

i. Identify the important hotel characteristics that influence hotel demand and measure them.
ii. Estimate how these hotel characteristics influence demand.
iii. Improve local search for hotels by incorporating the economic impacts of the hotel characteristics.

Specifically, in the first stage, we determine the particular hotel characteristics that are most highly valued by customers and thus contribute to the aggregate room prices of the hotels. Beyond the directly observable characteristics such as the "number of stars", provided by most third-party travel websites, many users also tend to value location characteristics such as proximity to the beach, or proximity to downtown, shopping areas etc. In our work, we incorporate the satellite image classification and use both human and computer intelligence (in the form of tagging and text mining), thereby leading

to a more comprehensive dataset. In the second stage, we use demand estimation techniques (BLP 1995, Berry and Pakes 2007, Song 2008) and estimate the economic value associated with each hotel characteristic. This enables us to quantitatively analyze how each feature influences demand and estimate its importance relative to other features. In the third stage, after inferring the economic significance of the location and service-based hotel characteristics, we incorporate them into designing a local ranking function. By doing so, we provide customers with the "best-value" hotels early on, hence improving the quality of local search for such hotels. In contrast to the existing research in recommender system which gives recommendations using a machine learning-based "black-box" style, this structural model-based approach tries to understand and capture the overall decision-making process of consumers.

2. Prior Literature

Our paper draws from multiple streams of work. A 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 needs to answer the following questions:

1. How can we identify which hotel attributes are most valued by consumers?
2. How can we automatically extract information about hotel attributes expressed in a product review?
3. How can we incorporate extracted sentiments and textual variables in a structural demand estimation model?

With the rapid growth and popularity of user-generated content on the Web, a new area of research applying text mining techniques 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 (Hu & Liu 2004, Pang & Lee 2004, Das &Chen 2007). More recent work has suggested that sentiment classification of consumer reviews is complicated, since consumers may provide a mixed review by praising some aspects of a product but criticizing other aspects. This stimulated additional research on identifying product features on which consumers expressed their opinions (Hu & Liu 2004, Scaffidi et al. 2007, Snyder & Barzilay 2007). Automated extraction of product attributes has also received attention in the recent marketing literature. In particular (Lee & 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.

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 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 underlying evaluations and does not demonstrate the importance of the underlying feature in the consumers' choice process. Is a hotel *"close to downtown"* more important to a business traveler than a hotel *"next to a beach"*? If so, then how much more important is it in influencing the traveler's booking decision? While questions of this nature might seem fuzzy, they can gain meaning if evaluated in the economic context surrounding consumer reviews and sales.

Our work is also related to models of demand estimation. One model that has made a significant impact in the applied econometrics field over the past decade is the random coefficient logit model, or BLP (Berry et al. 1995). However, as Berry and Pakes (2007) recently pointed out, BLP has its limitations. Specifically, it is shown to be problematic in welfare analysis. With the assumption of the product-level idiosyncratic logit "taste shock," the welfare numbers derived using BLP tend to heavily depend on this error term (Petrin 2002). As a logit model, it also suffers from the limitation of substitution patterns. This can be especially problematic when studying a market with a large number of products.

Due to the weaknesses of the product-level "taste shock" in logit models, a new model based on pure product characteristics has been proposed recently (Berry & Pakes 2007). The pure characteristic model (hereafter, PCM) differs from the BLP model in the sense that it does not contain the product-level "taste shock." It describes the consumer heterogeneity purely based on their different tastes towards individual product characteristic, without considerations on the tastes of certain product as a whole (i.e., brand preference). In this way, it eliminates the problems that BLP and other logit models suffer due to the presence of the product-level random error term. However, it is an ideal case. For what is observed in reality, the product-level idiosyncratic "tastes" from different consumers do exist in many markets. As pointed out in Song (2008), whether or not one should introduce the product-level "taste shock" should depend on the context of the market. Some markets are more likely to "benefit" from this shock, while others may not. In our study, we consider this "taste shock" from both the *product-level* (as in BLP) and the *product characteristic-level* (as in PCM). We discuss the model in details in the following sections.

Keeping in mind two levels of consumer heterogeneities introduced by (1) different travel contexts (i.e., family trip or business trip) and (2) different hotel characteristics, we propose a two-step random coefficient structural model to identify the latent weight distribution consumers assign to each hotel characteristic. The final outcome of our analysis allows us to compute the consumer surplus for each hotel. Based on this, in the third stage, we aim to generate a novel ranking approach which will provide customers with the “best-value" hotels early on, and thereby improve the quality of local search for such hotels.

The rest of the paper is organized as follows. The next section discusses the work related to the data preparation, including methods to identify the important hotel characteristics and the text mining techniques use to parse user-generated reviews. Then, in the following three sections, we provide an overview of our econometric approach, and how we will apply our results into a real world business application such as a ranking system for hotel search in a given location. Finally, in the last section we conclude with a summary of potential insights and some information on timelines for our deliverables.

3.Data Description

In this section, we discuss the data preparation work that is required. Some of it has already been undertaken while some others remain to be done. Our work leverages three types of user-generated content data:

- On-demand user generated content through Amazon Mechanical Turk
- Location descriptions based on social geotagging websites and image classification
- Service descriptions based on consumer reviews

We first introduce how we leverage Amazon Mechanical Turk to collect user preferences. Once we identify the preferences of the consumers, we then use other forms of user-generated content to understand the characteristics of the location, and the characteristics of the hotel, and how these are being taken into account by consumers.

3.1 Extraction of Hotel Characteristics using Amazon Mechanical Turk

Our analysis first requires knowing what aspects of a hotel are important for consumers, as these are the ones that ultimately determine the aggregate prices of the hotels. To perform the survey, we decided to rely on a “human-powered computing” technique, and use a semiautomatic human intelligence approach instead of a fully automated approach. Specifically, we used the Amazon

Mechanical Turk[2] (MTurk) service. AMT is an online marketplace, used to automate the execution of micro-tasks that require human intervention (i.e., cannot be fully automated using data mining tools). Task requesters post simple micro-tasks known as *HITs* (human intelligence tasks) in the marketplace. Workers browse the posted micro-tasks 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. Each user receives a small monetary compensation for completing the task.

There is a lot of evidence that users who contribute content on AMT are very representative of the general population. Specifically, the population of users that participate as workers on Mechanical Turk are mainly US residents, with an income and education distribution similar to the general population of online users. 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 non-expert users of Mechanical Turk can generate results of comparable quality as those generated by experts, especially after gathering results for the same micro task 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 non-noisy data. For our survey, we conducted a small pilot study. We asked 100 anonymous MTurk users for hotel characteristics that would influence their choice of a hotel. Our analysis identified two broad categories of hotel characteristics:

1. Location-based hotel characteristics (such as "Near the beach," "Near the waterfront,", "Near public transportation", "near shopping areas" and so on)
2. Service-based hotel characteristics (such as "Hotel class", "Quality of service", "Internal amenities" and so on)

Next, we describe how we use user-generated content to collect information about variables that are either too difficult or expensive to collect (e.g., density of shops around the hotel), or are subjective and based on consumer opinions (e.g., "quality of service").

[2] http://www.mturk.com

3.2 Extraction of Location Characteristics using Social Geotagging

For the location-based characteristics, we plan to combine user-generated content with automatic techniques, to be able to scale our data collection and generate data sets that are comprehensive at the national and even international level (i.e., tens or even hundreds of thousands of hotels). A first, automatic approach is to use a service like the Microsoft Virtual Earth Interactive SDK, which allows us to compute characteristics like "near restaurants and shops": using the automatic API we can perform automatically such "local search" queries.

However, a characteristic like "Near the beach" or "Near the park", or "Near Downtown" cannot be answered by existing mapping services. To measure such characteristics, we use a combination of social *geotagging*, in combination with automatic image classification of satellite images. The concept of geotagging has been popularized lately by photo sharing websites, in which users annotate their photos with the exact longitude and latitude. The concept however, has been extended and is now used in "wiki"-style websites, where users annotate maps with various types of annotations. For example, in the site Geonames.org, users annotate maps with tags like "beach," "bridge," "lake", "park," "school" and other similar tags. Such tagging allows us to generate a rich description of the location around each hotel, using features that are not available through existing mapping services. However, no matter how comprehensive the tagging is, there will always be locations that are not tagged. Therefore, we need ways to leverage the tag database, and allow automatic tagging of areas that lack any tags. For this, we use automatic image classification techniques together with satellite imagery, to automatically tag locations with tags that have significance for hotel customers.

3.3 Extraction of Location Characteristics using Image Processing

Image Data Retrieving: Consider for example the case where we are trying to understand whether a hotel is located in a downtown area, or next to a beach. (As a reminder, it is not possible to get this information from a mapping service, or from the TripAdvisor website.) For this, we extracted hybrid satellite images (sized 256 × 256 pixels) using the Visual Earth Tile System[3], for each of the (thousands) of hotel venues located in the United States, with 4 different zoom levels for each. These 4 x 9463 images were then used to extract information about the surroundings of the hotel, through image classification and through human inspection using Mechanical Turk.

[3] http://msdn2.microsoft.com/en-us/library/bb259689.aspx

Image Classification: In our work, to automatically tag satellite images, we first need to train our classification model. As a "training set" we plan to use the areas that have been already tagged by users of the social networking website. It has been shown in prior work that non-parametric classifiers, such as Neural Network, Decision Tree and Support Vector Machines (SVM) provide better results than parametric classifiers in complex landscapes (Lu and Weng 2007). Therefore, we tested various non-parametric classification techniques: (i) Decision Trees, which are widely used for training and classification of remotely sensed image data, (due to its capability to generate human interpretable decision rules and its relatively fast speed in training and classification), and (ii) Support Vector Machines (SVM), which are highly accurate and perform well for a wide variety of classification tasks (Fukuda and Hirosawa 2001).

In our preliminary experiments, we performed a small study to examine the performance of the classifier out-of-sample data. To perform the classification, we classified the out of sample images using Mechanical Turk; our results show that our SVM classifier has an accuracy of 91.2% for "Beach" image classification and 80.7% for "Downtown" image classification. We also used the C4.5 algorithm for classification, and noticed an accuracy increase for "Beach" and a decrease for "Downtown".

3.4 Extraction of Service Characteristics using Consumer Reviews

Service-based characteristics are used for specifying the performance of a hotel accommodation, including hotel amenities, appearance, service, and so on. There are 2 broad characteristics in this category: hotel class and internal amenities. Here, "hotel class" is an internationally accepted standard ranging from 1-5 stars representing low to high hotel grades. "internal amenities" is the aggregation of hotel internal amenities, including "24 hour front desk," "ice machine," "beautiful furnishings," "credit card payment," "cable TV," "pets allowed," "size of the room," "wheelchair accessible," "friendly staff," "free breakfast," "cleanliness," "wakeup call service," "nonsmoking," "gym," "iron," "internet reservation available," "high speed internet," "kids friendly service," "laundry services," "swimming pool," "parking," "kitchenette" and "spa."

This category contains characteristics related to the consumer review information, which is a broad option covering the "word of mouth" that a hotel has received online through popular travel sites, such as TripAdvisor or Travelocity. For example, we measured "word of mouth" by using the total number of reviews and the overall reviewer rating. Meanwhile, it has been widely aware that the actual textual contents of reviews play an important role in affecting products sales. In our case, instead of extracting

the polarity of the reviews, which can be more directly observed using simple numeric ratings, we look into two text style features, “subjectivity” and “readability.” Both of them are proved to be helpful and informative for consumers to make purchase decisions (Ghose and Ipeirotis 2008).

Furthermore, previous research suggested that identity information about reviewers in an online community shapes community members' judgment of products. Hence, the prevalence of reviewer disclosure of identity information is associated with changes in subsequent online product sales (Forman et al. 2008). Therefore, we decide to include one particular characteristic capturing the level of reviewers’ disclosure of their identity information – “real name or location.” Specifically, this binary characteristic describes whether or not a reviewer has revealed his/her real name or location information on their profile webpage. In summary, there are totally 5 broad types of characteristics in this category: *total number of reviews, overall reviewer rating, review subjectivity, review readability,* and *disclosure of reviewer identity information.* In order to capture more objectively the review text style, we decide to use a multiple-item method for *subjectivity* and *readability*. We include 2 sub-features for *subjectivity* and 5 sub-features for *readability*, each of which measures the review text style from an individual and independent point of view. For these sub-features, we will discuss in more details in the next subsection.

3.5 Extraction of Review Opinions Using Text Mining Methods

After identifying the important hotel characteristics, we now discuss how we effectively collect the corresponding data. Our dataset contains a total of 2117 hotels in the US. We collected data from comprehensive sources to conduct our study (see Table 1 for details). We have a 3 month hotel transaction data from Travelocity from November 2008 to January 2009, which contains the average transaction price per room per night and total number of rooms sold per night.

With regard to the service-based hotel characteristics, we extracted them from the website of TripAdvisor using fully automated JavaScript parsing engines. Since hotel amenities are not directly listed on TripAdvisor website, we retrieved them by following the link provided on the hotel web page, which randomly directs to one of its cooperative partner websites (i.e., Travelocity.com, Orbitz.com, Expedia.com, Priceline.com, Hotels.com, and etc.).

For the customer review-based characteristics, we collected the direct customer reviews through the websites of Travelocity as the travel agency itself. Meanwhile, to consider the indirect influences of online “word of mouth,” we also collected reviews from a third party - the Tripadvisor website, which is regarded as the world’s largest online travel community. The online reviews and reviewers’ information were collected on a daily basis up to January 2009. As for the text features of the reviews, we used the

method suggested by Ghose and Ipeirotis (2008). Specifically, in order to decide the probability of subjectivity for review text contents, we trained a classifier using as "objective" documents the hotel descriptions of each of the 2117 hotels in our data set, and we randomly retrieved 1000 reviews to construct the "subjective" examples of the training set. We conducted the training process by using a 4-grams Dynamic Language Model classifier provided by the LingPipe toolkit. With this in hand, we were able to acquire a subjectivity confidence score for each sentence in a review, thereby deriving the mean and standard deviation of this score, which represent the probability of the review being subjective. As for the readability, we looked into the cognitive cost for people to read the review contents. In particular, for each hotel, we considered all its up-to-date reviews to examine the average number of characters per review, average number of syllables per review, average number of spelling errors per review, average length of sentence as a complexity measurement (total number of characters divided by total number of sentences), and SMOG index as a difficulty-level measurement which indicates the number of years of formal education that a person requires in order to easily understand the text on the first reading.

3.6 Description and Summary Statistics of Variables

First, for a better understanding of the variables in our setting, we provide the data sources, definitions and summary statistics of all variables in Tables 1 and 2.

Table 1. Summary of Different Methods for Extracting Hotel Characteristics

Category	Hotel Characteristics		Methods
Transaction Data	Transaction Price (per room per night) Number of Rooms sold (per night)		Travelocity
Service-based	Hotel Class Hotel Amenities		TripAdvisor
Review-based	Number of Customer Reviews Overall Reviewer Rating Disclosure of Reviewer Identity Information		Travelocity and TripAdvisor
	Subjectivity	Mean Probability Std. Dev. of Probability	Text Analysis
	Readability	Number of Characters Number of Syllables Number of Spelling Errors Average Length of Sentence SMOG Index	
Location-based	Near the Beach Near Downtown		Image Classification
	External Amenities (Number of restaurants/ shopping destinations) Number of Local Competitors		Virtual Earth Interactive SDK
	Near the Interstate Highway Near the Lake/River Near Public Transportation		MTurk
	City Annual Crime Rate		FBI online statistics

Table 2. Definitions and Summary Statistics of Variables

Variable	Definition	Data Points	Mean	Std. Dev.	Range
PRICE	Transaction price per room per night	12651	126.54	79.50	12.00-978.00
CHARACTERS	Average number of characters	12651	412.81	401.34	0-2187
COMPLEXITY	Average sentence length	12651	39.26	36.66	0-164.42
SYLLABLES	Average number of syllables	12651	132.20	128.58	0-700
SMOG	SMOG index	12651	5.34	4.96	0-19.80
SPELLERR	Average number of spelling errors	12651	.59	.61	0-3.33
SUB	Subjectivity - probability mean	12651	.53	.49	0-1
SUBDEV	Subjectivity - probability std. dev	12651	.10	.15	0-.66
ID	Disclosure of reviewer identity	12651	.22	.41	{0,1}
CLASS	Hotel class	12651	2.58	1.37	1-5
COMPETITOR	Number of local competitors within 2 miles	12651	1.77	2.80	0-20
CRIME	City annual crime rate	12651	194.00	123.65	0-1310
AMENITYCNT	Total number of hotel amenities	12651	12.00	7.75	0-23
EXT	Number of external amenities within 1 mile, i.e., restaurants or shops	12651	4.95	7.37	0-27
BEACH	Beachfront within 0.6 miles	12651	.24	.43	{0,1}
LAKE	Lake or river within 0.6 miles	12651	.23	.42	{0,1}
TRANS	Public transportation within 0.6 miles	12651	.11	.31	{0,1}
HIGHWAY	Highway exits within 0.6 miles	12651	.68	.47	{0,1}
DOWNTOWN	Downtown area within 0.6 miles	12651	.69	.46	{0,1}
REVIEWCNT	Total number of reviews	12651	16.86	157.22	0-202
RATING	Overall reviewer rating	12651	.91	1.61	0-5

4.Structural Model

In this section, we discuss how we develop our two-step random coefficient structural model and describe how we apply it to empirically estimate the distribution of consumer preferences towards different hotel characteristics in our setting.

4.1 Two-step Random Coefficient Model

We would like keep two questions in mind as we specify the model: Is this market influenced by product-level taste shocks? If it is, then where does this shock come from? For the first question, the answer is "yes" . As for the second question, instead of using a "brand" specific taste shock as most of the current studies found in the other markets, we propose to introduce a "taste shock" originating from consumers' different *travel contexts* in the hotel industry. Specifically, we define a consumer's purchase decision making behavior in the hotel market to be in accordance with the following two-step procedure.

In the first step, the consumer is going to find a subset of hotels which are evaluated by the online travel communities to provide the best expertise in the travel context that matches her own. For instance, if a consumer wants to go on a business trip, she would only be interested in a subset of hotels which are better specialized in business services; while if a consumer plans to take her four-year kid for a family fun trip, she would be more likely to look for those hotels which are evaluated to be kids friendly. For this purpose, we classified each hotel as belonging to one of the following eight types of ``travel category" : *Family Trip*, *Business Trip*, *Romantic Trip*, *Tourists Trip, Trip with Kids*, *Trip with Seniors*, *Pets Friendly* and *Disabilities Friendly*. This classification was based on the distributions of traveler evaluation and demographics from the online reviews on Travelocity.com and Tripadvisor.com. In order to capture heterogeneity in consumers' travel context, we introduce an idiosyncratic "taste shock" at this step. This is similar in flavor to the product-level "taste shock" in the BLP (1995) model.

Then, in the second step, once the consumer has picked a specific travel category, she will obtain a corresponding subset of hotels which satisfy her travel requirement. From this subset, she can make her further decision purely based on her evaluation of the quality of the hotels. In this case, we use the pure characteristic model to capture the vertical differentiation among hotels within the same category.

This model can be written in the following form:

$$u_{ij^k t} = X_{j^k t}\beta_i - \alpha_i P_{j^k t} + \xi_{j^k t} + \varepsilon_{it}^k, \qquad (1)$$

where, *i* represents a consumer, j^k represents hotel *j* with category type *k* ($1 \le k \le 7$), and *t* represents a hotel market which in our case is defined as a "city-night" combination. In this model, β_i and α_i are random coefficients that capture consumers' heterogeneous tastes towards different observed hotel characteristics, *X*, and towards the average price per night, *P*, **respectively**. ξ represents the set of hotel characteristics that are unobservable to the econometrician.

Notice that ε_{it}^{k} with a superscript *k* represents a travel context level "taste shock". This idiosyncratic "taste shock" only appears in the first step when consumers decide to choose a certain travel category type *k*. However, it disappears thereafter, because ε_{it}^{k} will remain consistent within each *k* at the second step. This ε would have been otherwise written with a superscript *j* corresponding to the hotel level "taste shock" as in the BLP model and all other Logit models, or simply dropped as in the pure characteristics model (PCM). Thus, by using the two-step model, the utility $u_{ij^{k}t}$ for consumer *i* from choosing hotel *j* with category type *k* in market *t* can be represented as shown in equation (1).

Due to the computational complexity and data restriction, estimating a unique set of weights for each consumer is near impossible. In order to make this model tractable, we make some further assumptions about β_i and α_i. One is to assume that these weights are distributed among consumers per some known statistical distribution, i.e., $\beta_i \sim (\beta_i \mid \overline{\beta}, \sigma_\beta)$ and $\alpha_i \sim (\alpha_i \mid \overline{\alpha}, \sigma_\alpha)$. Our goal is then to estimate the means ($\overline{\beta}$, $\overline{\alpha}$) and the standard deviations (σ_β, σ_α) of these two distributions. The means correspond to the set of coefficients on hotel characteristics and on hotel price, which measure the average weight placed by consumers; while the standard deviations provide a measure of the consumer heterogeneity in those weights. This assumption provides an analytical solution to our problem.

Furthermore, we notice these heterogeneities result from some particular demographic attributes of consumers. For example, the variance in the price elasticity is very likely to be a result of different incomes of different consumers. Therefore, we make further assumptions about the standard deviations: $\sigma_\alpha \sim \alpha_I I_i$, where I_i represents the income whose distribution can be learned from the consumer demographics; $\sigma_\beta \sim \beta_v v_i$, where $v_i \sim N(0,1)$ represents some random factor that will influences people's preferences towards individual hotel characteristics.

Therefore, we rewrite our model in this following form:

$$u_{ij^{k}t} = \delta_{j^{k}t} + X_{j^{k}t}\beta_v v_i - \alpha_I I_i P_{j^{k}t} + \varepsilon_{it}^{k}, \qquad (2)$$

where $\delta_{j^{k}t} = X_{j^{k}t}\overline{\beta} - \overline{\alpha}P_{j^{k}t} + \xi_{j^{k}t}$, represents the mean utility of hotel *j* with category type *k* in market *t*. β_v and α_I are the set of parameters to be estimated. Note that for computational complexity reason, we assume β_v and α_I to be both scalars in our setting.

4.2 Estimation

With the model in hand, now we discuss how we identify the values for the parameters. As mentioned in the previous subsection, our goal here is to estimate the mean and variance of β_i and α_i. We apply methods similar to those used in Berry and Pakes (2007) and Song (2008). In general, with a given starting value of $\theta_0 = (\alpha_I^0, \beta_v^0)$, we look for the mean utility δ such that the model predicted market share equates the observed market share. From there, we form a GMM objective function using the moment conditions that the mean of unobserved characteristics is uncorrelated with instrumental variables. Based on this, we identify a new value of $\theta_1 = (\alpha_I^1, \beta_v^1)$, which will be used as the starting point for the next round iteration. This procedure is repeated until the algorithm finds the optimal value of θ that minimizes the GMM objective function. More specifically, we conduct the estimation in the following three stages.

(1) Calculating Market Shares

- **Market share for hotel conditional on travel category type.**

We start with computing the market share for each hotel within a particular travel category type *k*. This can be done in two steps. First, we consider our model to be a vertical model condition on v_i. By doing so, we are able to integrate out one dimension of customer heterogeneity. Then, we integrate over the distribution of v_i to get the total market share.

More specifically, we begin by ordering the hotels based on their price in an ascending fashion. Thus, consumer *i* chooses hotel j^k if and only if its utility exceeds the utility from any of the other hotels with the same travel category type:

$$\delta_{j^k t} + X_{j^k t}\beta_v v_i - \alpha_I I_i P_{j^k t} > \delta_{h^k t} + X_{h^k t}\beta_v v_i - \alpha_I I_i P_{h^k t}, \quad \forall h^k \in S_k \text{ and } h^k \neq j^k,$$

where S_k represents the subset of hotels with expert type *k*.

This can be transformed to

$$(\delta_{j^k t} - \delta_{h^k t}) + (X_{j^k t} - X_{h^k t})\beta_v v_i > \alpha_I I_i (P_{j^k t} - P_{h^k t}).$$

Therefore, conditioning on v_i, a consumer with income type I_i will choose hotel j^k if and only if

$$I_i < \min_{j^k > h^k} \frac{(\delta_{j^k t} - \delta_{h^k t}) + (X_{j^k t} - X_{h^k t})\beta_v v_i}{\alpha_I (P_{j^k t} - P_{h^k t})} \equiv \overline{\Delta}(\square \mid \theta, v),$$

$$\text{and } I_i > \max_{j^k < h^k} \frac{(\delta_{j^k t} - \delta_{h^k t}) + (X_{j^k t} - X_{h^k t})\beta_v v_i}{\alpha_I (P_{j^k t} - P_{h^k t})} \equiv \underline{\Delta}(\square \mid \theta, v).$$

Let $F(\square)$ denote the cdf of I_i, and $G(\square)$ denote the cdf of v_i. The market share of hotel *j* with expert type *k* can be calculated as the following:

$$s_{j^k} = \int [F(\bar{\Delta}_{j^k}(\square|\theta, v)) - F(\underline{\Delta}_{j^k}(\square|\theta, v))] \, 1\,[\bar{\Delta}_{j^k}(\square|\theta, v) > \underline{\Delta}_{j^k}(\square|\theta, v)] dG(v) \tag{3}$$

where $1\,[\square]$ is the indicator function for the condition, and θ is a vector containing parameters α_I and β_v. Note here, in order to compute the income upper bound $\bar{\Delta}(\square|\theta, v)$ and lower bound $\underline{\Delta}(\square|\theta, v)$, we need the value of θ. However, θ is unknown at this time. Therefore, we choose to use an iteration method as suggested by previous studies, where we start from an initial point (α_I^0, β_v^0). We will discuss more estimation details on this in the second stage.

Nevertheless, given the set of values for θ, this integration is typically not analytic. For this reason, we use a Monte Carlo simulation to approximate it. Since v_i follows the standard normal distribution $v_i \sim N(0,1)$, we can obtain an unbiased estimator of this integral by taking ns_v random draws of v_i from $N(0,1)$ as follows

$$s_{j^k}(\delta, p, X; \theta, F, G_{ns}) \equiv \frac{1}{ns_v} \sum_{i}^{ns} [F(\bar{\Delta}_{j^k}(\square, \theta, v_i)) - F(\underline{\Delta}_{j^k}(\square, \theta, v_i))] \, 1\,[\bar{\Delta}_{j^k}(\square, \theta, v_i) > \underline{\Delta}_{j^k}(\square, \theta, v_i)] \tag{4}$$

- **Market share for each travel category type.**

Now, let's look at the market share for a particular expert type *k*. A consumer *i* chooses expert type *k* if and only if the best hotel (which provides the highest utility) within this expert type exceeds the best hotels within any other expert types:

$$\max_{j^k \in S_k}(\delta_{j^k t} + X_{j^k t}\beta_v v_i - \alpha_I I_i P_{j^k t}) + \varepsilon_{it}^k > \max_{j^r \in S_r}(\delta_{j^r t} + X_{j^r t}\beta_v v_i - \alpha_I I_i P_{j^r t}) + \varepsilon_{it}^r, \quad \forall r \neq k.$$

Therefore, similar as in the Logit models, by assuming ε with a type I extreme value distribution, we can calculate the market share for an expert type *k* as the following

$$s_k = \iint \frac{\exp(\max_{j^k \in S_k}(\delta_{j^k t} + X_{j^k t}\beta_v v_i - \alpha_I I_i P_{j^k t}))}{\sum_{r=1}^{K} \exp(\max_{j^r \in S_r}(\delta_{j^r t} + X_{j^r t}\beta_v v_i - \alpha_I I_i P_{j^r t}))} f(I) g(v) dI dv. \tag{5}$$

- **Final market share for hotel *j* with travel category type *k*.**

Based on the discussion above, the probability for hotel *j* with expert type *k* to be chosen can be calculated as:

$$s_{j^k} = \iint_{I_i, v_i \in C_{j^k}} \frac{\exp(\delta_{j^k_t} + X_{j^k_t}\beta_v v_i - \alpha_I I_i P_{j^k_t})}{\sum_{r=1}^{K} \exp(\max_{j^r \in S_r}(\delta_{j^r_t} + X_{j^r_t}\beta_v v_i - \alpha_I I_i P_{j^r_t}))} f(I)g(v)dIdv. \quad (6)$$

In this equation, $I_i, v_i \in C_{j^k}$ represents the set of consumers who choose hotel *j* with expert type *k*. Because we have two dimensions of heterogeneities, this market share equation contains two levels of integrals. We rewrite this by decomposing the inner integral into two parts as suggested in Song (2008):

$$s_{j^k} = \int_{v_i \in C_{j^k}} \left([F(\bar{\Delta}_{j^k}(\square, \theta, v_i)) - F(\underline{\Delta}_{j^k}(\square, \theta, v_i))] \times \int_{I_i \in C_{j^k}} \frac{\exp(\delta_{j^k_t} + X_{j^k_t}\beta_v v_i - \alpha_I I_i P_{j^k_t})}{\sum_{r=1}^{K} \exp(\max_{j^r \in S_r}(\delta_{j^r_t} + X_{j^r_t}\beta_v v_i - \alpha_I I_i P_{j^r_t}))} h(I)dI \right) g(v)dv, \quad (7)$$

where we extract the part of $[F(\bar{\Delta}_{j^k}(\square, \theta, v_i)) - F(\underline{\Delta}_{j^k}(\square, \theta, v_i))]$ out of the inner integral, and substitute $h(I)$ for $f(I)$, with $h(I) = \frac{f(I)}{F(\bar{\Delta}_{j^k}(\square, \theta, v_i)) - F(\underline{\Delta}_{j^k}(\square, \theta, v_i))}$. Again, these integrals are not analytic, but we can use a Monte Carlo simulation-based approach to approximate their values based on the distributions $G(v)$ and $h(I)$:

$$s_{j^k} = \frac{1}{ns_v} \sum_{v_i \in C_{j^k}} \left([F(\bar{\Delta}_{j^k}(\square, \theta, v_i)) - F(\underline{\Delta}_{j^k}(\square, \theta, v_i))] \frac{1}{ns_I} \sum_{I_i \in C_{j^k}} \frac{\exp(\delta_{j^k_t} + X_{j^k_t}\beta_v v_i - \alpha_I I_i P_{j^k_t})}{\sum_{r=1}^{K} \exp(\max_{j^r \in S_r}(\delta_{j^r_t} + X_{j^r_t}\beta_v v_i - \alpha_I I_i P_{j^r_t}))} \right). \quad (8)$$

(2) Solving Mean Utility δ

With the market share being derived, we can then identify the mean utility δ by equating the estimated market share to the observed market share conditioning on a given $\theta = (\alpha_I, \beta_v)$. As we can see, this problem can be essentially reduced to a procedure of solving a system of nonlinear equations. In our case, there are $\sum_{k=1}^{K} J^k$ nonlinear equations (where J^k is the total number of hotels within expert type *k*) and $\sum_{k=1}^{K} J^k$ unknown variables (δ being a $\sum_{k=1}^{K} J^k$ dimension vector).

To find a solution, we applied the contraction mapping method suggested by Berry et al. (1995) in BLP estimation. In practice, this approach found us the closest solution for our settings and the iteration procedure provided a very close form to locate the roots rapidly and stably. In order to test the robustness of the results, we also tried different initial values of δ in the iteration. The final solution is proved to be consistent.

(3) Solving α_I and β_v

Considering the endogeneity of price, we use a GMM estimator and form an objective function by interacting the unobservable ξ with a set of instrumental variables. In our case, we use the average price of the "same-star rating" hotels in the same market as an instrument for price. By minimizing the GMM objective function, we can find a proper set of α_I and β_v. Set the new α_I and β_v as the starting points to recalculate the market share in step (1) and solve for the new mean utility in step (2). This whole procedure continues to iterate until the algorithm finds the optimal combination of α_I, β_v and δ.

5. Empirical Analysis and Results

One thing to note here is that the above dataset contains hotels from Travelocity which may or may not have online customer reviews. Since one important goal of our study is to examine the potential economic value from the online user generated content, we therefore narrow down the sample to consist of those hotels that have at least one review from either Travelocity or TripAdvisor website. This leads to a smaller dataset of 1479 hotels. The estimation results from this filtered dataset (I) and from the unfiltered dataset (II) are both shown in Table 3.

Note that the coefficients of a large majority of variables are statistically significant from both datasets. "Price" presents a negative sign, which is consistent with the "law of demand" in reality. The higher the price, the lower the quantity demanded. Moreover, based on the estimated signs of the coefficients, we can qualitatively analyze the economic impacts of different hotel characteristics.

There are at least four location-based characteristics which have a positive impact on hotel demand: "Hotel external amenities," "Public transportation," "Highway", and "Downtown." These characteristics strongly imply that the location and geographical convenience for a hotel can make a big difference in attracting consumers. Hotels providing easy access to public transportation (such as subway or bus stations), highway exits, restaurants and shops, or to downtown area, can have much higher demand from consumers. [4]

[4] Note that "Downtown" appears to be insignificant in (I), but it turns out to be highly significant at the level of 0.1% in (II). The similar finding also applies to "Amenity count".

Table 3. Estimation Results

Variable	Coef. (Std. Err)[I]	Coef. (Std. Err)[II]
	Means	
Price	-.1768*** (*.0289*)	-.0080 (.0144)
CHARACTERS	.0155*** (.0020)	.0108*** (.0015)
COMPLEXITY	-.0121*** (.0026)	-.0070*** (.0020)
SYLLABLES	-.0482*** (.0063)	-.0331*** (.0048)
SMOG	.1137*** (.0280)	.0650*** (.0195)
SPELLERR	-.1575*** (.0416)	-.1250*** (.0318)
SUB	-.8268* (.3322)	-.2265† (.1317)
SUBDEV	-.2298** (.0758)	-.2221*** (.0576)
ID	.1366*** (.0270)	.1044*** (.0172)
CLASS	.0421*** (.0128)	-.0049 (.0055)
COMPETITOR	-.0853*** (.0118)	-.1435*** (.0147)
CRIME	-.1523*** (.0174)	-.0598*** (.0095)
AMENITYCNT	.0022 (.0020)	.0023* (.0010)
EXT	.0066*** (.0019)	.0052*** (.0011)
BEACH	.0693* (.0335)	.1035*** (.0178)
LAKE	-.1452*** (.0289)	-.1214*** (.0164)
TRANS	.1495*** (.0290)	.00003** (9.61e-06)
HIGHWAY	.1332*** (.0272)	.0848*** (.0153)
DOWNTOWN	.0275 (.0287)	.0713*** (.0160)
REVIEWCNT	-.0890*** (.0099)	-.0736*** (.0067)
RATING	.0504*** (.0082)	.0202*** (.0061)
Constant	2.1245*** (.4047)	.2609** (.0822)
	Standard Deviations	
α_I	.000001	.000012
β_v	.001	.002
GMM Obj Value	2.476e-17	8.445e-17

*** Significant at a 0.1% level.

** Significant at a 1% level.

* Significant at a 5% level.

† Significant at a 10% level.

I Estimation based on the filtered dataset.

II Estimation based on the unfiltered dataset.

Three location-based characteristics have a negative impact on hotel demand. Not surprisingly, one of them is the “Annual Crime Rate.” The higher the average crime rate reported in a local area, the lower the desirability of consumers for staying in a hotel located in that area. This indicates that neighborhood safety usually plays an important role in hotel industry. Another factor that has a negative impact is the number of ``Local competitors” within 2 miles. This is reasonable because given the fact that the total market shares in a certain area often remain consistent during a three-month period, an increase in the number of competitors within the same hotel market will usually lead to a decrease in an individual hotel’s market share. Consumers have more choices when they look for a hotel, which causes the demand for each individual hotel to drop on an average. The third location-based characteristic that shows a negative impact is “Lake/River.” This is quite interesting because most times people would prefer to choose a hotel near a lake or by the river side. However, after examining our data more carefully, we found that our transaction period happened to be the coldest three months in a year. Due to the freezing season, it is very likely that lake or river front become less desirable for travelers. For comparison purposes, “Beach”, however, shows a positive impact. This is because most beach sites in our dataset are located in the south where the weather is warm even in winter. Therefore, the desirability of a “walkable” beachfront is not negatively influenced by the winter season.

For Service-based characteristics, we notice both “Class” and “Amenity count” pose a positive influence on hotel quality and thus increasing the demand. This agrees with what is observed in reality. Hotels with more amenities and higher star level usually are better preferred by consumers.

For Review-based characteristics, in general, we found that “Overall reviewer rating” has a positive impact, while “Total number of reviews”, however, has a negative one. The total number of reviews causes the hotel demand to drop. This implies the high possibility that the majority of reviews may present a negative attitude and provide a comparably low rating from the reviewers. In order to examine this, we further look into our data and found that the mean of rating is indeed very low (1.29 for the filtered dataset and 0.91 for the unfiltered one, out of 5), which to a large extent supports our estimation results.

Another important factor from the customer reviews is their text style. According to the estimation results, we found all the readability and subjectivity characteristics are statistically significant to hotel demand from both datasets. Among all the readability sub-features, “Complexity”, “Syllables” and “Spelling Errors” have a negative sign, which shows that the average length of a sentence, the total number of syllables and spelling errors in a review will all have a negative impact on hotel sales. This implies that most consumers would prefer to read reviews with shorter sentences, less syllables and

fewer spelling errors in total. Thus, hotels with such reviews usually could attract more consumers for this reason. On the other hand, variables “Characters” and “SMOG index” present a positive influence. This implies that consumers also appreciate longer reviews with more characters in total, and with a more professional writing style.

For the subjectivity sub-features, both “Mean Subjectivity” and “Subjectivity standard deviation” turn out to be negative. From a consumer point of view, online reviews for hotels are highly favored to contain more objective information (such as descriptions for room quality, location, neighborhood environment, etc.). This gives us an important implication that hotel is an experience good. Hence, its quality is difficult to observe in advance but can be ascertained upon consumption. Therefore, when a consumer looks for a hotel, she would strongly prefer to obtain as much objective information as possible from the previous consumers’ experiences. For a comparison, previous study found a positive impact of the subjectivity probability mean on the video-audio player and computer sales (Ghose and Ipeirotis 2008). This is most likely because these goods are both search goods which qualities are much easier to observe beforehand. Therefore, in those cases, people would want to know more about the personal opinions and feelings from previous consumers as complementary information for their final decisions. For the “Probability standard deviation”, our finding is consistent with the previous study, which implies people’s preference towards a “consistent subjectivity style” from online customer reviews. The last but not least review-based characteristic is “Disclosure of reviewer identity”. It shows a strong positive impact on hotel demand. This result strengthens the findings from previous work (Forman et al. 2008), which suggests that the identity information about reviewers in the online travel community can indeed shape community members' judgment towards a certain hotel, and the prevalence of reviewer disclosure of identity information has a strong positive influence on the subsequent hotel sales.

Robustness Check s

In order to consider the potential influence of user-generated content from other online communities outside Travelocity, we conducted the same estimation based on additional reviews collected from a third party - the Tripadvisor website, which is regarded as the world’s largest online travel community. The results are qualitatively consistent with our findings here. Table 4 shows the estimation results using both the internal reviews from Travelocity and the external reviews from TripAdvisor.

Table 4. Estimation Results Using Third Party Reviews

Variable	Coef. (Std. Err)[I]	Coef. (Std. Err)[II]
	Means	
Price	-.1251*** (*.0294*)	.0166 (.0145)
CHARACTERS	.0133*** (.0020)	.0087*** (.0015)
COMPLEXITY	-.0114*** (.0026)	-.00004 (.0009)
SYLLABLES	-.0411*** (.0063)	-.0264*** (.0047)
SMOG	.1308*** (.0278)	-.0109 (.0102)
SPELLERR	-.1061* (.0417)	-.0903** (.0318)
SUB	-.2799*** (.0791)	-.0176 (.0551)
SUBDEV	-.1941* (.0916)	-.1376* (.0703)
ID	.1106*** (.0269)	.1006*** (.0171)
CLASS	.0364*** (.0103)	-.0103† (.0055)
COMPETITOR	-.0518*** (.0121)	-.0436*** (.0067)
CRIME	-.1445*** (.0172)	-.0460*** (.0095)
AMENITYCNT	.0039* (.0020)	.0079*** (.0011)
EXT	.0068*** (.0019)	.0057*** (.0011)
BEACH	.1236*** (.0337)	.1423*** (.0179)
LAKE	-.1106*** (.0289)	-.0984*** (.0164)
TRANS	.1489*** (.0289)	.1774*** (0227)
HIGHWAY	.1045*** (.0271)	.0610*** (.0153)
DOWNTOWN	.0510† (.0285)	.0719*** (.0160)
TL_REVIEWCNT	-.0868*** (.0113)	-.0771*** (.0072)
TA_REVIEWCNT	-.1432*** (.0099)	-.0580*** (.0053)
TL_RATING	.0250** (.0087)	.0108† (.0064)
TA_RATING	.0502*** (.0094)	.0135* (.0060)
Constant	1.2942*** (.2459)	.1690 (.1631)
	Standard Deviations	
α_I	.000001	.000002
β_v	.001	.002
GMM Obj Value	2.129e-16	3.424e-17

*** Significant at 0.1% level.

** Significant at 1% level.

* Significant at 5% level.

† Significant at 10% level.

I Estimation based on the filtered dataset.

II Estimation based on the unfiltered dataset.

6. Consumer Surplus-Based Hotel Ranking

After we have estimated the parameters in the model and interpreted the underlining economic value of the hotel characteristics, we can derive the consumer surplus from our model. From there, we propose a new ranking approach for hotels based on the consumer surplus. As we discussed in the previous sections, in order to capture the consumer heterogeneity, we represent the excess utility from each hotel for each consumer as consisting of two parts: the mean and the standard deviation. The mean utility provides us a good estimation of how much consumers in general can benefit from choosing this particular hotel, and the standard deviation of utility describes the variance of this benefits from different consumers. In our case, we are interested to know what the excess utility, or consumer surplus, is for consumers on an aggregate level to choose a certain hotel. Therefore, we define the consumer surplus from hotel *j* with expert type *k* as the sum of its mean excess utility $\bar{\mu}_{ij^k t}$ divided by the mean price elasticity $\bar{\alpha}$ over all markets:

$$CS_{j^k} = \sum_t \frac{1}{\bar{\alpha}} \bar{\mu}_{ij^k t} \tag{9}$$

6.1 Ranking Hotels Based on Consumer Surplus

We thereby propose a new ranking approach for hotels based on the consumer surplus of each hotel for consumers on an aggregate level. This ranking idea is based on how much "extra value" consumers can obtain after paying for that hotel, which is what consumers really care about. If a hotel provides a comparably higher surplus for consumers on an aggregate level, then it should appear on the top part of our ranking list. The higher ranked hotels can provide consumers with higher surplus value, thus should be more recommended to consumers.

Since the mean price elasticity is consistent over all hotels, and our final goal is to *relatively* compare hotels and rank them based on the consumer surplus, we can simply ignore $\bar{\alpha}$, which gives us the following form:

$$CS_{j^k} = \sum_t \frac{1}{\bar{\alpha}} \bar{\mu}_{ij^k t} \sim \sum_t \bar{\mu}_{ij^k t} = \sum_t \delta_{j^k t} \tag{10}$$

Therefore, we can rank the hotels by their mean utility δ, which represents a reasonable estimate of their surplus. After estimating the economic impact for each hotel characteristic, we propose to design a local hotel ranking function based on the *consumer surplus* estimation. Then, we rank all the hotels according to their "value for the money" in a descending order, which gives a best valuation on the hotel

cost performance and provides customers with the best valued hotels consequently. A preliminary ranking result for New York City is listed in Appendix A.

6.2 Experimental Evaluation of Our Ranking

To evaluate the quality of our ranking technique, we conducted a user study using Amazon Mechanical Turk (AMT). First, we generated different rankings for the top-20 hotels, in various areas, according to a set of baseline criteria: price low to high, price high to low, maximum online review count, hotel class, hotel size (number of rooms), and number of internal amenities. We then computed the consumer surplus for each hotel, and ranked the hotels in each city according to their surplus. Then, we performed blind tests, presenting various lists to 100 anonymous AMT users and asking them which ranking list they prefer. Further, we asked users to compare pairs of lists and tell us which of the hotel ranking lists they prefer the most. We tested the results for a few large cities like New York city, and the results were highly encouraging. A large majority of customers preferred our ranking when listed side-by-side with the other competing baseline techniques ($p = 0.05$, sign test).

We also asked consumers why they chose a particular ranking, to understand better how users interpret the surplus-based ranking. In our NYC experiment, the majority of the users indicated that our consumer surplus-based ranking provides hotels with better locations that can best capture the featured interests of the city. Meanwhile, people favored our ranking in the sense that it promotes the idea that price is not the only main factor in rating the quality of hotels. Instead, a good ranking recommendation should be able to satisfy customers' multidimensional preferences. Moreover, users strongly preferred the diversity of the returned results given that the list consisted of a mix of hotels cutting across several price and class ranges. In contrast, the other ranking approaches tend to list hotels of only one type (e.g., very expensive hotels). We found that a ranking system generated based on consumer surplus returns a better variety of hotels, covering 10% 5-star, 35% 4-star, 35% 3-star, 10% 2-star and 10% 1-star hotels in a given city. It generally starts out with lower class hotels and increases to 5-star hotels, providing a logical way to present the information on the screen which will help customers in their decision-making procedure. Based on the qualitative opinions of the users, it appears that diversity in hotel choices is indeed an important factor that improves the satisfaction of consumers, and an economic approach for ranking introduces diversity naturally. This result seems intuitive: if a specific segment of the market systematically appeared to be underpriced, then market forces would move the prices for the whole segment accordingly. However, this effect may be less pronounced with individual hotels, especially under a personalized consumer surplus calculation.

7. Conclusion and Implications

In this paper, we empirically estimate the economic value of different hotel characteristics, especially the location-based and review-based characteristics given the associated local infrastructure. We propose a two-step random coefficient structural model taking into consideration of two-level consumer heterogeneities introduced by different travel contexts and different hotel characteristics. Combining this state-of-the-art econometric model with user-generated contents data using techniques from text mining, image classification, on-demand annotations and geo-information system tools, we are able to examine a unique dataset consisting of totally 12,651 observations for 2117 different hotels located in the United States for 3 months and infer the economic significance of hotel characteristics from that. Based on this, we qualitatively interpret the economic impacts of various hotel characteristics and incorporate them into a novel ranking solution using the derived consumer surplus. By doing so, we are able to provide customers with the “best-value" hotels early on, hence improving the quality of local search for such hotels. Meanwhile, for the hotel business owners, our analyses will also strengthen their ability to make more accurate pricing decisions, thereby achieving better revenue management in the short-term and the long-term.

On a broader note, the objective of this paper is to show how user generated content (UGC) on the Internet can be incorporated in a demand estimation model and provide insights for using text mining techniques in economics and marketing research. Simultaneously, such research can also highlight the value of using an economic context to computer scientists to estimate both the intensity and the polarity of UGC, especially in reviews and blogs. Towards this, we empirically estimate the economic value of different hotel characteristics, including both service based and location-based characteristics from multiple sources of UGC. Our research allows us to not only quantify the economic impact of hotel characteristics, but also by reversing the logic of this analysis, allows us to identify the characteristics that most influence the demand for a particular hotel. After inferring the economic significance of each characteristic, we qualitatively interpret the economic impacts of various hotel characteristics and we incorporate the economic value of hotels characteristics into a local ranking function. By doing so, we hope to be able to improve the quality of travel search engines on the Internet

References

- Archak, N., Ghose, A., Ipeirotis, G. 2008. Deriving the pricing power of product features by mining consumer reviews, Working Paper, SSRN.
- Berndt, E. 1996. The Practice of Econometrics: Classic and Contemporary, Addison-Wesley Publishing Company, Reading, MA.
- Berry, S. 1994. Estimating Discrete Choice Models of Product Differentiation. *RAND Journal of Economics* (25). pp. 242-262.
- Berry, S., Levinsohn, J., and Pakes, A. 1995. Automobile Prices in Market Equilibrium. *Econometrica* (63). pp. 841-890.
- Berry, S. and Waldfogel, J. 2001. Do Mergers Increase Product Variety? Evidence from Radio Broadcasting. *Quarterly Journal of Economics* (116). pp. 969-1007
- Berry, S., Linton, O., and Pakes, A. 2004. Limit Theorems for Estimating the Parameters of Differentiated Product Demand Systems. *Review of Economic Studies* (71). pp. 613-654.
- Berry, S. and Pakes, A. 2007. The Pure Characteristics Demand Model. *International Economic Review* (48). pp. 1193-1225.
- Bickart, B., R. Schindler. 2001. Internet forums as influential sources of consumer information. Journal of Interactive Marketing 15(3) 31–40.
- Bowen, J. T., Sparks, B. A. 1998. Hospitality marketing research: A content analysis and implications for future research, *International Journal of Hospitality Management* (17:2), pp. 125-144.
- Bull, A. 1998. The Effects of Location and Other Attributes on the Price of Products Which are Place-sensitive in Demand, PhD thesis, Griffith University, 1998.
- Chen, Y., J. Xie. 2004. Online consumer reviews: A new element of marketing communications mix. Working Paper. SSRN.
- Chevalier, J., D. Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *J. Marketing Res*. 43(3) 345-354.
- Clemons, E., G. Gao, L. Hitt. 2006. When Online Review Meets Hyperdifferentiation: a Study of Craft Beer Industry. *Journal of Management Information Systems, 23(2), 149-171.*
- Das, S., M. Chen. 2007. Yahoo! for Amazon: Sentiment extraction from small talk on the web. Management Science 53(9) 1375–1388.
- Dellarocas, C., N. Awad, M. Zhang. 2007. Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *J. Interactive Marketing* 21(4) 23-45.
- Duan, W., B. Gu, A. B. Whinston. 2008. Do online reviews matter? An empirical investigation of panel data. *Decision Support Systems* 45(4) 1007-1016.
- 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.
- Franklin, S., Wulder, M. 2002. Remote sensing methods in medium spatial resolution satellite data land cover classification of large areas, *Progress in Physical Geography* (26), pp. 173-205.

- Fukuda, S., Hirosawa, H. 2001. Support vector machine classification of land cover: application to polarimetric SAR data, in *Proceedings of the IEEE International Geoscience & Remote Sensing Symposium*, pp. 187-189.
- Garbers, J., Niemann, M., and Mochol, M. 2006. A Personalized Hotel Selection Engine, in *Proceedings of the third European Semantic Web Conference*, Budva, Montenegro, June.
- 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.
- Ghose, A., P. Ipeirotis, A. Sundararajan. 2006. The Dimensions of Reputation in Electronic Markets. Working paper, New York University, NY.
- Ghose, A., P. Ipeirotis. 2008. Estimating the socio-economic impact of product reviews: Mining text and reviewer characteristics. Working paper, New York University.
- Ghose, A. 2009. Internet exchanges for used goods: An empirical analysis of trade patterns and adverse selection. *MIS Quarterly* 33(1) 1-30.
- Godes, D., D. Mayzlin. 2004. Using online conversations to study word of mouth communication. *Marketing Sci.* 23(4) 545-560.
- Hu, M., B. Liu. 2004. Mining and summarizing customer reviews. Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004). 168–177.
- Hu, N., Pavlou, P. A., Zhang, J. 2006. Can online reviews reveal a product's true quality? Empirical findings and analytical modeling of online word-of-mouth communication. *Proceedings of the seventh ACM conference on electronic commerce*, pp. 324-330.
- Lancaster, K. 1966. A new approach to consumer theory. *Journal of Political Economy* (74), pp. 132-157.
- Lancaster, K. 1971. Consumer Demand: A New Approach. Columbia University Press, NY.
- Lee, T., E. Bradlow. 2007. Automatic construction of conjoint attributes and levels from online customer reviews. University of Pennsylvania, The Wharton School Working Paper OPIM WP 06-08-01.
- Li, X., L. Hitt. 2008. Self selection and information role of online product reviews1. Information Systems Research 19(4).
- Liu, B., M. Hu, J. Cheng. 2005. Opinion observer: Analyzing and comparing opinions on the Web. Proceedings of the 14th International World Wide Web Conference (WWW 2005). 342–351.
- McFadden, D. 1973. Conditional Logit Analysis of Quantitative Choice Behavior. *Frontiers in Econometrics*, ed. By P. Zarambka. New York: Academic Press.
- McFadden, D. 2001. Economic Choices. *The American Economic Review* 91(3). pp. 351-37.
- Nevo, A., 2000. A Practitioner's Guide to Estimation of Random Coefficients Logit Models of Demand, *Journal of Economics & Management Strategy*, 9(4), 513-548.
- Nevo, A., A. Rosen 2008. Identification with Imperfect Instruments, Working Paper, SSRN.
- Pang, B., Lee, L. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pp. 271-278.

- Petrin, A. 2002. Quantifying the benefits of new products: The case of the minivan, *Journal of Political Economy.*
- Rosen, S. 1974. Hedonic prices and implicit markets: Product differentiation in pure competition, *Journal of Political Economy* (82:1), pp. 34-55.
- Scaffidi, C., K. Bierhoff, E. Chang, M.Felker, H. Ng, C. Jin. 2007. Red opal: Product-feature scoring from reviews. Proceedings of the 8th ACM conference on Electronic commerce (EC'07). 182–191.
- Sheng, V., F. Provost, P. Ipeirotis. 2008. Get another label? Improving data quality and data mining using multiple, noisy labelers. Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2008).
- Snow, R., B. O. Connor, D. Jurafsky, A. Ng. 2008. Cheap and fast but is it good? evaluating non-expert annotations for natural language tasks. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008).
- Song, M. 2008. A Hybrid Discrete Choice Model of Differentiated Product Demand with an Application to Personal Computers." Simon School Working Paper No. FR 08-09. Available at SSRN: http://ssrn.com/abstract=1315271.
- White, P. 1998. Site and situation determinants of hotel room rates, Master's thesis.

Appendix A Hotel Ranking Result in New York City

Table 6. Top 20 New York City Hotels

Hotel	Price($)	Consumer Surplus($)
1. W New York – Union Square	368.86	-4.4803
2. The Michelangelo	425.24	-5.0790
3. Fitzopatrick Grand Central Hotel	286.51	-5.6683
4. Affinia 50	310.73	-5.7198
5. W New York – Times Square	364.19	-5.8926
6. Wall Street Inn	239.40	-5.9241
7. Washington Square Hotel	213.00	-6.9896
8. Shelburne Murray Hill	238.48	-7.1017
9. Hilton New York	248.67	-7.4999
10. Roger Williams Hotel	274.93	-7.5051
11. Millenium Hilton	255.98	-7.5324
12. Hotel Mela	242.69	-7.7747
13. Hampton Inn Manhattan Chelsea	192.06	-7.9626
14. The Muse	398.78	-8.2040
15. Holiday Inn Express Madison Square	196.32	-8.2720
16. The Shoreham	249.72	-8.3237
17. The New York Helmsley	272.42	-8.3281
18. Club Quarters Downtown	179.54	-8.4428
19. Grand Hyatt new York	314.62	-8.4695
20. New York Palace Hotel	353.83	-8.5902